Banjercito

Diplomado en Ciencia de Datos

# Tarea 502

## Módulo V — Semana 1

Alan Badillo Salas

Noviembre 2024

## Introducción

El Diplomado en Ciencia de Datos ha llegado a su quinto módulo titulado "*Deep Learning*". En este módulo revisaremos a profundidad las redes neuronales generales, recurrentes y convolutivas para resolver problemas generales y particulares de forma automática, mejorando así la predicción y pronóstico en los análisis y casos de estudio relacionados.

En la semana 1 hemos hecho un repaso sobre los temas más importantes de los módulos anteriores, partiendo del uso de Python, Numpy y Pandas, así como la probabilidad y estadística bayesiana y de los modelos de Regresión y Clasificación de Machine Learning. También se introdujo el concepto de *Perceptrón* y su aplicación para predecir un objetivo a través de sus características.

En esta tarea se reforzarán las habilidades para hacer un análisis probabilístico y estadístico con Excel y Python.

# Tarea 502 — Probabilidad y Estadística de Deuda

Una financiera posee una lista de 30 créditos no completados que fueron otorgados a clientes de diferente edad, escolaridad, años de servicio y si eran casados. La financiera requiere entender algunas probabilidades y estadísticos descriptivos de los datos.

Genera una hoja de Excel que contenga los siguientes datos sobre los 30 créditos otorgados y la deuda total de cada cliente:

| Cliente | Crédito | Deuda | Edad | Escolaridad | Años | Casado |
|---|---|---|---|---|---|---|
| 1 | $5,000.00 | $800.00 | 25 | Preparatoria | 1 | 0 |
| 2 | $6,000.00 | $200.00 | 25 | Licenciatura | 2 | 0 |
| 3 | $8,000.00 | $3,500.00 | 27 | Preparatoria | 1 | 0 |
| 4 | $9,000.00 | $3,000.00 | 28 | Licenciatura | 3 | 1 |
| 5 | $10,000.00 | $2,500.00 | 29 | Preparatoria | 2 | 0 |
| 6 | $12,000.00 | $6,000.00 | 26 | Preparatoria | 1 | 1 |
| 7 | $12,000.00 | $3,000.00 | 24 | Preparatoria | 3 | 0 |
| 8 | $13,000.00 | $2,500.00 | 26 | Licenciatura | 2 | 0 |
| 9 | $15,000.00 | $4,000.00 | 31 | Preparatoria | 1 | 0 |
| 10 | $15,000.00 | $500.00 | 33 | Licenciatura | 2 | 1 |
| 11 | $16,000.00 | $3,000.00 | 34 | Licenciatura | 2 | 0 |
| 12 | $17,000.00 | $1,500.00 | 40 | Licenciatura | 5 | 0 |
| 13 | $19,000.00 | $800.00 | 36 | Licenciatura | 5 | 1 |
| 14 | $19,000.00 | $1,200.00 | 34 | Maestría | 3 | 0 |
| 15 | $20,000.00 | $900.00 | 42 | Licenciatura | 7 | 1 |
| 16 | $23,000.00 | $6,000.00 | 48 | Maestría | 5 | 0 |
| 17 | $23,000.00 | $13,000.00 | 52 | Licenciatura | 12 | 0 |
| 18 | $23,000.00 | $4,000.00 | 54 | Licenciatura | 5 | 1 |
| 19 | $25,000.00 | $5,000.00 | 36 | Maestría | 8 | 0 |
| 20 | $25,000.00 | $9,000.00 | 39 | Licenciatura | 5 | 0 |
| 21 | $26,000.00 | $8,000.00 | 41 | Licenciatura | 2 | 1 |
| 22 | $27,000.00 | $14,000.00 | 56 | Licenciatura | 24 | 0 |
| 23 | $28,000.00 | $6,000.00 | 54 | Maestría | 5 | 1 |
| 24 | $28,000.00 | $8,000.00 | 28 | Licenciatura | 3 | 1 |
| 25 | $30,000.00 | $6,000.00 | 31 | Licenciatura | 6 | 1 |
| 26 | $34,000.00 | $7,000.00 | 58 | Licenciatura | 12 | 0 |
| 27 | $34,000.00 | $8,000.00 | 64 | Preparatoria | 24 | 1 |
| 28 | $35,000.00 | $8,000.00 | 66 | Maestría | 40 | 0 |
| 29 | $35,000.00 | $6,000.00 | 56 | Preparatoria | 36 | 1 |
| 30 | $35,000.00 | $12,000.00 | 58 | Licenciatura | 28 | 1 |

Table 1: Datos de clientes

Calcula los conteos siguientes:

- Número total de clientes que están casados
- Número de clientes menores o iguales a 30 años
- Número de clientes entre 31 y 39 años
- Número de clientes entre 40 y 55 años
- Número de clientes mayores o iguales a 56 años
- Número de clientes con Preparatoria
- Número de clientes con Licenciatura
- Número de clientes con Maestría
- Número de clientes con 1 o 2 años de servicio
- Número de clientes con 3 a 5 años de servicio
- Número de clientes con 6 o más años de servicio

Calcula las siguientes probabilidades (porcentaje de que ocurra):

- Probabilidad que un cliente esté casado
- Probabilidad que un cliente tenga hasta 30 años
- Probabilidad que un cliente tenga entre 31 y 39 años
- Probabilidad que un cliente tenga entre 40 y 55 años
- Probabilidad que un cliente tenga al menos 56 años
- Probabilidad que un cliente tenga Preparatoria
- Probabilidad que un cliente tenga Licenciatura
- Probabilidad que un cliente tenga Maestría
- Probabilidad que un cliente tenga 1 o 2 años de servicio
- Probabilidad que un cliente tenga de 3 a 5 años de servicio
- Probabilidad que un cliente tenga 6 o más años de servicio

Calcula los siguientes estadísticos:

- Suma de la deuda en crétidos entre $5,000 y $10,000
- Suma de la deuda en crétidos entre $11,000 y $15,000

- Suma de la deuda en crétidos entre $16,000 y $20,000
- Suma de la deuda en crétidos entre $21,000 y $25,000
- Suma de la deuda en crétidos entre $26,000 y $30,000
- Suma de la deuda en crétidos entre $31,000 y $35,000
- La deuda menor en crétidos entre $5,000 y $10,000
- La deuda menor en crétidos entre $11,000 y $15,000
- La deuda menor en crétidos entre $16,000 y $20,000
- La deuda menor en crétidos entre $21,000 y $25,000
- La deuda menor en crétidos entre $26,000 y $30,000
- La deuda menor en crétidos entre $31,000 y $35,000
- La deuda mayor en crétidos entre $5,000 y $10,000
- La deuda mayor en crétidos entre $11,000 y $15,000
- La deuda mayor en crétidos entre $16,000 y $20,000
- La deuda mayor en crétidos entre $21,000 y $25,000
- La deuda mayor en crétidos entre $26,000 y $30,000
- La deuda mayor en crétidos entre $31,000 y $35,000
- Promedio de la deuda en crétidos entre $5,000 y $10,000
- Promedio de la deuda en crétidos entre $11,000 y $15,000
- Promedio de la deuda en crétidos entre $16,000 y $20,000
- Promedio de la deuda en crétidos entre $21,000 y $25,000
- Promedio de la deuda en crétidos entre $26,000 y $30,000
- Promedio de la deuda en crétidos entre $31,000 y $35,000
- Desviación estándar de la deuda en crétidos entre $5,000 y $10,000
- Desviación estándar de la deuda en crétidos entre $11,000 y $15,000
- Desviación estándar de la deuda en crétidos entre $16,000 y $20,000
- Desviación estándar de la deuda en crétidos entre $21,000 y $25,000
- Desviación estándar de la deuda en crétidos entre $26,000 y $30,000
- Desviación estándar de la deuda en crétidos entre $31,000 y $35,000

Calcula las siguientes probabilidades conjuntas:

- Probabilidad que un cliente tenga hasta 30 años y esté casado
- Probabilidad que un cliente tenga hasta 30 años y tenga Preparatoria
- Probabilidad que un cliente tenga hasta 30 años y tenga Licenciatura
- Probabilidad que un cliente tenga hasta 30 años y tenga Maestría
- Probabilidad que un cliente tenga entre 31 y 39 años y esté casado
- Probabilidad que un cliente tenga entre 31 y 39 años y tenga Preparatoria
- Probabilidad que un cliente tenga entre 31 y 39 años y tenga Licenciatura
- Probabilidad que un cliente tenga entre 31 y 39 años y tenga Maestría
- Probabilidad que un cliente tenga entre 40 y 55 años y esté casado
- Probabilidad que un cliente tenga entre 40 y 55 años y tenga Preparatoria
- Probabilidad que un cliente tenga entre 40 y 55 años y tenga Licenciatura
- Probabilidad que un cliente tenga entre 40 y 55 años y tenga Maestría
- Probabilidad que un cliente tenga al menos 56 años y esté casado
- Probabilidad que un cliente tenga al menos 56 años y tenga Preparatoria
- Probabilidad que un cliente tenga al menos 56 años y tenga Licenciatura
- Probabilidad que un cliente tenga al menos 56 años y tenga Maestría

Calcula las siguientes probabilidades condicionales:

- Probabilidad que un cliente casado tenga hasta 30 años
- Probabilidad que un cliente casado tenga entre 31 y 39 años
- Probabilidad que un cliente casado tenga entre 40 y 55 años
- Probabilidad que un cliente casado tenga al menos 56 años
- Probabilidad que un cliente casado tenga Preparatoria
- Probabilidad que un cliente casado tenga Licenciatura
- Probabilidad que un cliente casado tenga Maestría
- Probabilidad que un cliente con Preparatoria esté casado
- Probabilidad que un cliente con Licenciatura esté casado
- Probabilidad que un cliente con Maestría esté casado

Calcula las siguientes probabilidades condicionales sobre la deuda relativa, es decir, la deuda entre el crédito, por ejemplo, para una deuda de $800, y un crédito de $5,000 la deuda relativa es $800/5000 = 0.16 = 16\%$:

- Cliente casado que tenga una deuda relativa de hasta 20%
- Cliente casado que tenga una deuda relativa de entre 20% y 40%
- Cliente casado que tenga una deuda relativa de entre 40% y 60%
- Cliente casado que tenga una deuda relativa de entre 60% y 80%
- Cliente casado que tenga una deuda relativa mayor a 80%
- Cliente con Preparatoria que tenga una deuda relativa de hasta 20%
- Cliente con Preparatoria que tenga una deuda relativa de entre 20% y 40%
- Cliente con Preparatoria que tenga una deuda relativa de entre 40% y 60%
- Cliente con Preparatoria que tenga una deuda relativa de entre 60% y 80%
- Cliente con Preparatoria que tenga una deuda relativa mayor a 80%
- Cliente con Licenciatura que tenga una deuda relativa de hasta 20%
- Cliente con Licenciatura que tenga una deuda relativa de entre 20% y 40%
- Cliente con Licenciatura que tenga una deuda relativa de entre 40% y 60%
- Cliente con Licenciatura que tenga una deuda relativa de entre 60% y 80%
- Cliente con Licenciatura que tenga una deuda relativa mayor a 80%
- Cliente con Maestría que tenga una deuda relativa de hasta 20%
- Cliente con Maestría que tenga una deuda relativa de entre 20% y 40%
- Cliente con Maestría que tenga una deuda relativa de entre 40% y 60%
- Cliente con Maestría que tenga una deuda relativa de entre 60% y 80%
- Cliente con Maestría que tenga una deuda relativa mayor a 80%

Sigue los pasos del *script* para generar distintos reportes sobre las probabilidades conjuntas de estar casado por escolaridad y el porcentaje de deuda relativa por escolaridad.

Carga los datos de los créditos con Pandas:

```
import pandas
creditos = pandas.read_excel("/content/p502.xlsx")
creditos
```

**Pregunta**: ¿Cuál es la deuda máxima de los clientes con Preparatoria?

Calula la columna de deuda relativa (deuda entre crédito):

```
creditos["Deuda Rel"] = 100 * \
    creditos["Deuda"] / creditos["Crédito"]
creditos
```

**Pregunta**: ¿Cuáles son las dos deudas relativas más grandes, qué escolaridad tienen y en qué rango de edad están?

Crea un DataFrame para reportar las probabilidades condicionales de que dado que el cliente tiene cierta escolaridad y esté o no casado:

```
import numpy
reporteCasadoEscolaridad = pandas.DataFrame(
    numpy.zeros((3, 2)),
    index=["Preparatoria", "Licenciatura", "Maestria"],
    columns=["Casado", "No casado"])
reporteCasadoEscolaridad
```

**Pregunta**: ¿Por qué se usa $numpy.zeros((3, 2))$?

Llena el DataFrame con los conteos usando dos filtros, el primero indicará cual es la escolaridad y el segundo si está casado, se usará la probabilidad condicional $P(B|A) = P(B, A)/P(A)$:

```
for escolaridad in ["Preparatoria", "Licenciatura", "Maestria"]:
  for casado, columna in [(1, "Casado"), (0, "No casado")]:
    filtro1 = creditos["Escolaridad"] == escolaridad
    filtro2 = creditos["Casado"] == casado
    proba = len(creditos[filtro1 & filtro2]) / \
        len(creditos[filtro1])
    reporteCasadoEscolaridad.loc[escolaridad, columna] = \
        proba * 100
```

**Pregunta**: ¿Cuál es la probabilidad de que un cliente que tiene maestría, no esté casado?

Grafica en un pastel para cada escolaridad la probabilidad de estar o no casado:

```
import matplotlib.pyplot as pyplot

for escolaridad in ["Preparatoria", "Licenciatura", "Maestria"]:
  filtro1 = creditos["Escolaridad"] == escolaridad
  filtro2A = creditos["Casado"] == 1
  filtro2B = creditos["Casado"] == 0
  A = len(creditos[filtro1 & filtro2A])
  B = len(creditos[filtro1 & filtro2B])
  pyplot.pie([A, B],
             labels=[f"Casado ({A})", f"No casado ({B})"],
             autopct="%.1f%%")
  pyplot.title(f"Cliente con {escolaridad}")
  pyplot.savefig(f"r1-{escolaridad}-casado.png")
  pyplot.show()
```

En la Figura 1 observamos las gráficas de pastel que comparar la probabilidad de estar casado por escolaridad.

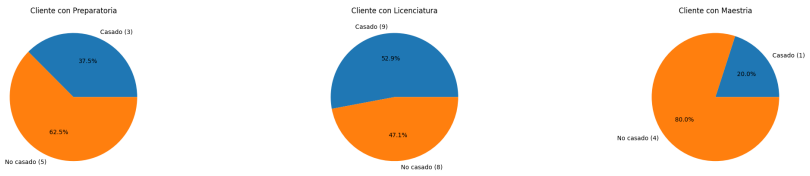


Figura 1: Probabilidad de estar casado en clientes por escolaridad

**Pregunta**: ¿Qué escolaridad tienen mayor probabilidad de que un cliente no esté casado y cuál de que si lo esté?

Haz un análisis similar para la probabilidad condicional por escolaridad pero ahora para el porcentaje de deuda relativa:

```
import numpy

reporteDeudaEscolaridad = pandas.DataFrame(numpy.zeros((3, 6)),
    index=["Preparatoria", "Licenciatura", "Maestria"],
    columns=["10%", "20%", "30%", "40%", "50%", "60%"])

for escolaridad in ["Preparatoria", "Licenciatura", "Maestria"]:
    for lim_inf, lim_sup, columna in [(-1, 10, "10%"),
        (10, 20, "20%"), (20, 30, "30%"),
        (30, 40, "40%"), (40, 50, "50%"), (50, 100, "60%")]:
      filtro1 = creditos["Escolaridad"] == escolaridad
```

```
    filtro2 = creditos["Deuda Rel"] > lim_inf
    filtro3 = creditos["Deuda Rel"] <= lim_sup
    proba = len(creditos[filtro1 & filtro2 & filtro3]) / \
      len(creditos[filtro1])
    reporteDeudaEscolaridad.loc[escolaridad, columna] = \
      proba * 100

reporteDeudaEscolaridad
```

**Pregunta**: ¿Cuál es la probabilidad de tener una deuda entre 20% y 30% dado que el cliente tiene Maestría y cuál dado que el cliente tiene Preparatoria?

Genera una gráfica de mapa de calor sobre los datos del reporte (escolaridad y porcentaje de deuda relativa):

```
import seaborn
seaborn.heatmap(reporteDeudaEscolaridad)
pyplot.savefig("escolaridad_deuda_heatmap.png")
```

En la Figura 2 se muestra el mapa de calor de la probabilidad de tener cierto porcentaje de deuda relativa por escolaridad:

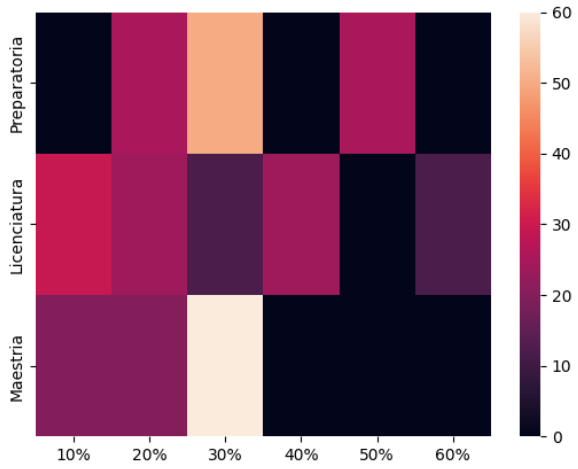


Figura 2: Mapa de calor con la escolaridad contra el porcentaje de deuda

**Pregunta**: ¿Cómo se interpreta esta gráfica, dónde concentra el mayor porcentaje de deduda relativa cada escolaridad?

Finalmente, crea una representación en gráfica de barras de los mismos datos:

```
import matplotlib.pyplot as pyplot

pyplot.bar(
    numpy.arange(len(reporteDeudaEscolaridad.columns)) + 0,
    reporteDeudaEscolaridad.loc["Preparatoria"],
    label="Preparatoria",
    width=0.2)
pyplot.bar(
    numpy.arange(len(reporteDeudaEscolaridad.columns)) + 0.2,
    reporteDeudaEscolaridad.loc["Licenciatura"],
    label="Licenciatura",
    width=0.2)
pyplot.bar(
    numpy.arange(len(reporteDeudaEscolaridad.columns)) + 0.4,
    reporteDeudaEscolaridad.loc["Maestria"],
    label="Maestria",
    width=0.2)

pyplot.xticks(
    numpy.arange(len(reporteDeudaEscolaridad.columns)) + 0.2,
    reporteDeudaEscolaridad.columns)

pyplot.title("Deuda Relativa por Escolaridad")
pyplot.xlabel("Deuda Relativa")
pyplot.ylabel("% Créditos")
pyplot.legend()

pyplot.savefig("escolaridad_deuda_bar.png")

pyplot.show()
```

En la Figura 3 se muestra las barras agrupadas de la probabilidad de tener cierto porcentaje de deuda relativa por escolaridad:

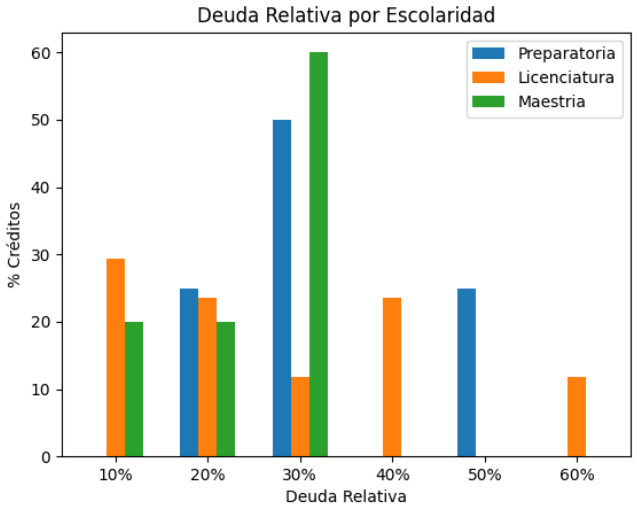


Figura 3: Barras agrupadas con la escolaridad contra el porcentaje de deuda

**Pregunta**: ¿Cómo se interpreta esta gráfica, dónde concentra el mayor porcentaje de deduda relativa cada escolaridad?

Escribe tus conclusiones y piensa en cómo se haría un análisis por edad en lugar de escolaridad.