**UNIVERSIDAD CATÓLICA BOLIVIANA”SAN PABLO”  
MAESTRÍA EN CIENCICIA DE DATOS Versión 2  
MATERIA MINERIA DE DATOS 1**

**DATASET DE PRÁCTICA**

**DATASET CREDITO**

1. Quitar filas duplicadas
2. Quitar columnas duplicadas
3. Identificar columnas que tienen valores constantes
4. Identifique el porcentaje de valores faltantes en las columnas
5. Para el dataset hacer un análisis gráfico documentado de las variables:

* Histograms
* Quartiles
* Scatter plots
* Distributions
* Boxplots
* Bar plot
* Heatmaps

1. Aplicar expresión matemática para el cálculo (CreditAmount / Duración)\*residence\_since
2. Obtenga una muestra para cada conjunto de Training, Testing y Validation
3. Discretizar duración, credit\_amount, age, number\_of\_dependents.
4. Obtenga la matriz de correlación de las variables numéricas.
5. Obtenga la tabla de relación de frecuencias entre los atributos job y purpose.
6. Mapear a valores diferentes para atributos own\_telephone, foreign\_worker.
7. Normalizar el campo instalment\_commitment por frecuencia.
8. Normalizar el campo instalment\_commitment en el rango (-1,2)
9. Detectar outliers por medio de Inter-quartil range
10. Utilice One-hot encoding en campo purpose.
11. Obtenga las instancias que cumplan la condición de un monto de crédito mayor a 3000
12. Obtenga las instancias que cumplen con la condición de que job sea igual a “skilled” y que el monto de crédito sea mayor a 2000.
13. Obtenga suma total del camp credit\_amount agrupado por el campo purpose “new car”.