**UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”  
CARRERA DE INGENIERÍA DE SISTEMAS  
MAESTRÍA EN CIENCIA DE DATOS Tercera Versión  
MATERIA: Minería de Datos I  
Prof. MSc. Oswaldo Figueroa Domejean  
Febrero 2022**

**NOMBRE: RAMON WILDER SERDAN CARDENAS**

**EXAMEN Nro. 1**

1. Cual es la diferencia entre el training set, testing set y el validation set, en un dataset de un ejemplo que usted plantee.

El training set es la data en la cual toma lugar el entrenamiento, es decir, el aprendizaje.

El Testing Set proporciona información en relación a la precision de nuestro modelo después de haber completado la fase de entrenamiento.

El Validation set ayuda a mejorar el modelo y hacer correcciones (fine-tuning).

Ejemplo. Para predecir el nivel de Reservas Internacionales debemos separar nuestra información histórica para realizar el entrenamiento y aprendizaje de la variable de interés, posteriormente ejecutaremos el Test para conocer la precisión de pronóstico de nuestro modelo. La proporción de la data para la validación nos informará y proporcionará de elementos para hacer correcciones al modelo.

1. En un dataset se pueden encontrar valores atípicos (i.e. outliers), cuales son los criterios que usted plantearía en el procesamiento de este tipo de datos.

Desde la perspectiva económica, los outliers se pueden dar por variables o factores no tomados en cuenta en nuestro set de información. Por ejemplo, pandemia, decisiones políticas, etc. Desde este punto de vista, debemos fundamentar y explicar el surgimiento de dichos valores atípicos. Podemos separar la porción de información atípica para aplicar unos supuestos y un modelo diferente en comparación a los datos “normales”.

Asimismo, se podría utilizar una variable dummy para señalar que existe un cambio de tendencia anormal en un periodo de tiempo determinado o para ciertas observaciones.

1. Explique el propósito del proceso Exploratoty Data Analysis y plantee dos ejemplos en los cuales se debieran procesar las columnas (i.e. features) de un data set de ejemplo que usted plantee.

El objetivo del proceso EDA es obtener máxima información de un dataset descubriendo su estructura, extrayendo variables importantes, detectando valores atípicos, variables con alta correlación mutua, datos faltantes, etc.

Los ejemplos en los que se debe procesar features serían:

* Variables cualitativas susceptibles de convertirse con cuantitativas para un mejor tratamiento con las herramientas informáticas. Ejm. Fecha en literal. Estado civil.
* Standarización y normalización de variables de magnitud mayor (PIB de socios comerciales) o considerablemente menor a otras variables (tasa de interés)
* Aplicación de funciones a datos heterogénoses. Ejm. Aplicación de logaritmos a datos muy grandes en magnitud.
* Mediante un algoritmo KNN algunos features con datos faltantes pueden asumir el valor más próximo o de alta probabilidad de ocurrencia. Ejm. Tasa de inflación por departamento.

1. Porque la monotonicidad puede ser un problema en la preparación de datos y cuál sería la solución para contar con un archivo preparado para la ejecución de las técnicas de Mineria de Datos. Explique con un ejemplo.

Una monotonicidad de datos representa un problema debido a que no aporta información relevante para nuestro modelo y la frecuencia de la información que implica.

Correspondería no tomar en cuenta dicho feature o en algún caso la transformación de la variable ya que la magnitud no aportaría información relevante para explicar el modelo.

Un ejemplo es la id de una persona entrevistada en una encuesta.

1. La Minería de Datos permite identificar patrones en los datos analizados. Plantee aplicaciones de esta área del conocimiento en actividades propias del rubro de Organización donde usted trabaja o conoce.

En el caso de la variable Retiros de Bóveda, las técnicas de Minería de Datos coadyuvarían a identificar patrones de comportamiento de los agentes involucrados en el retiro de divisas de bóvedas del Banco Central de Bolivia (que afectan el nivel de Reservas Internacionales). Dichos agentes serían las entidades financieras, las instituciones públicas y el sector privado exportador e importador que demandan divisas para el normal desempeño de sus funciones. La información que se analizaría sería los estados financieros diarios de los agentes, el tipo de cambio real de los socios comerciales, las medidas económicas de política como la emisión de bonos, contratación de créditos externos y los precios de commodities.