**UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”  
FACULTAD DE INGENIERÍA  
CARRERA INGENIERÍA DE SISTEMAS  
MAESTRÍA EN CIENCIA DE DATOS TERCERA VERSIÓN  
MATERIA: MINERÍA DE DATOS I**

**NOMBRE: RAMON WILDER SERDAN CARDENAS**

**EXAMEN Nro. 2**

1. Plantee el proceso en el ámbito de modelos supervisados, comenzando desde que se cuenta con un dataset, la elaboración del training, testing y el validation por medio de un ejemplo concreto que usted plantee.

Los modelos supervisados se caracterizan por tener etiquetas que se transforman en el target a ser aprendido por el algoritmo.

El ejemplo, en el ámbito de mi profesión como economista, son las medidas de política a corto plazo a aplicar según la tendencia de variables relevantes que se constituirían en los features.

El target (class) del dataset es la información histórica de las decisiones de política monetaria expansiva o contractiva (1 o 0). Los features serían la tendencia de las Reservas Internacionales (pérdida a ganancia semanal), el tipo de cambio real (ITCER) y el flujo de caja de los Entidades de Intermediación Financiera.

Dicho dataset de medidas de corto plazo se dividiría en un, tradicional 80% para el training set y 20% para el validation y testing.

Ya que el target es una variable dicotómica podríamos aplicar una técnica Naive Bayes ya que se cuenta con registros de volumen considerables.

Una vez que el modelo haya aprendido las medidas a tomar dados los features históricos, se aplicaría como respaldo para la decisión de ser expansivos o contractivos en las políticas monetarias de corto plazo.

1. Plantee un ejemplo de su autoría en el cual se pueda utilizar la técnica del Naive Bayes, identificando claramente el objeto del proceso de análisis, como también los features y el target.

En el Banco Central se realizan aproximadamente 1200 registros contables diarios de instituciones públicas que solicitan mediante cartas físicas diferentes transacciones en sus cuentas. Dichas cartas tienen un formato estándar para cada institución, mediante el sistema de correspondencia, se cuenta con una copia electrónica en formato PDF.

Es poca la información necesaria para procesar contablemente una solicitud (features):

* Institución solicitante.
* número de cuenta origen
* número de cuenta destino
* departamento
* concepto
* moneda
* monto
* área encargada del procesamiento (target)

El problema es la tardanza en la clasificación de dichas solicitudes para el procesamiento contable que se realiza manualmente y para el registro estadístico que corresponde a un proceso más analítico.

Se plantea una solución utilizando la técnica de Naive Bayes que recoja la información de las solicitudes pasadas y considerando el monto (algunos montos pequeños de algunos departamentos son procesados con esquemas contables especiales), departamentos (según el concepto y el departamento algunas operaciones pueden ser rechazadas), moneda y concepto, pueda clasificar correctamente qué área del banco debe realizar dicho registro. Asimismo, la solución pueda sugerir una clasificación estadística de acuerdo a la información pasada.

1. Describa con ejemplos las razones por las cuales la marca NULL o NaN puede ser un problema que tergiverse la aplicación de técnicas de Ciencia de Datos en diversos ámbitos.

Los valores inexistente o no números pueden llevar a conclusiones irrelevantes y equivocadas.

Por ejemplo, si nos encontramos con dos variables “Desembolsos” y “amortizaciones” en nuestro dataset, lógicamente cuando un registro tiene un valor en el campo “desembolsos” no vamos a encontrar un valor en el campo “amortizaciones” puesto que una transacción o es desembolso o es amortización. Entonces, en el dataset veremos muchos valores nulos. Ahora, si el analista quiere interpretar una consulta o formula agregada de un solo campo sin tomar en cuenta el otro, puede realizar un llenado de datos faltantes con los valores mas frecuentes de manera totalmente errada.

Otro ejemplo, son las diferencias en campos agregados cuando se utilizan diferentes herramientas de análisis como Excel, Python, R, etc. Es frecuente ver que existen diferencias en los promedios y el número de ocurrencias en los resultados. En mi experiencia, es mejor utilizar Python cuando existen muchos registros y se tiene la certeza de la existencia de datos Nan y nulos, ya que Python tiene mejores herramientas en el tratamiento de estas marcas.

1. Describa por medio de ejemplos concretos de un dataset, cuáles son las razones por las que se realiza un análisis exploratorio de datos procesando los features para adecuarlos a las necesidades de las técnicas que se pudieran aplicar, utilice al menos tres tipos de features y sus transformaciones.

Discretización de una variable continua. Un feature, como el Ingreso, puede ser más útil cuando se lo analiza mediante niveles (bajo, medio, alto) y dependiendo de la técnica también se le puede aplicar un one\_hot\_enconding.

Variables Dummy. Algunos features como catástrofes naturales, medidas de política económica, inestabilidad electoral, etc. Son eventos que son importantes para el análisis y la manera de transformarlas sería asignándolas cero y unos para que sean aptas en una técnica de minería de datos.

Outliers. En economía, la variable Ingresos extraordinarios, son outliers que deben ser tomados en cuenta puesto que afectan a las conclusiones del análisis. Un ejemplo concreto es el financiamiento mediante bonos soberanos los cuales ocurren muy rara vez y elevan considerablemente el flujo y saldo de la deuda externa. Se debe aplicar una técnica de transformación de datos verificando previamente la existencia de dichos datos atípicos.

Missing data. Es común en la ciencia económica la poca disponibilidad de observaciones en datos relevantes. La transformación se la debe hacer tomando en cuenta los datos más frecuentes y sobre todo el expertise del analista para clasificar los datos faltantes.