ITESO

DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

PROYECTO DE APLICACIÓN MODULO 2-3

ESTE PROYECTO TIENE COMO FINALIDAD LA EVALUACIÓN DE LOS CONOCIMIENTOS ADQUIRIDOS DURANTE EL MODULO 2 Y 3.

INTRODUCCIÓN:

Actualmente, con los avances en las tecnologías de la información, la generación de datos de diversos tipos en un solo día tiene volúmenes muy altos y con tendencia creciente. Con el fin del aprovechamiento de la información valiosa que pueda estar oculta en los datos que se generan, se requieren tener conocimientos básicos de manejo de información y de análisis exploratorio de datos.

De forma general, a menos que la persona sea una experta en el fenómeno en el cual se están generando los datos, el ingeniero que se disponga al análisis de los datos generados debe de realizar un análisis exploratorio para rescatar las características básicas que poseen los datos. Además de realizar un agrupamiento de los mismos datos en base a una característica de interés.

Por otra parte, existen problemas en los cuales se tiene un conjunto de datos en donde se tienen identificados las clases de cada uno, y el problema consiste en reconocer datos que no estaban considerados anteriormente en las clases correctas. Esta clasificación de datos en clases se puede lograr con el diseño de modelos logísticos que utilicen un “subset” de datos donde se encuentren una muestra de los datos con sus respectivas clases de pertenencias.

Normalmente los datos no son separables por medio de una superficie lineal, por lo que se requiere utilizar una superficie no lineal. La utilización de un polinomio crea un nuevo conjunto de datos en un espacio de mayor dimensión en el que posiblemente si son linealmente separables. Esta transformación de los datos en otras dimensiones es llamada como cambio de kernel. Es decir, si los datos pasan por un polinomio, se dice que se usó un kernel polinomial.

Como se vio en el curso, tanto la regresión logística como el algoritmo de maquina vector soporte pueden ser usado para clasificar datos, y para ambos se requieren usar cambios de kernel generalmente.

OBJETIVO:

Consideren un conjunto de datos de la base de datos que se le asignó o eligió.

El objetivo de este proyecto de aplicación se puede separar en los siguientes puntos:

1. **Cargar la información de la base de datos proporcionada en el formato pertinente.**
2. **Realizar la limpieza y extracción de la información estadística básica que tienen los datos que se están analizando (Lo que se realizó en el proyecto anterior).**
3. **Realizar un estudio de calidad de los datos para ver el tipo de datos comentar sobre ellos.**
4. **Realización del análisis de componentes principales de la basa de datos a analizar, mostrar gráficas y decidir si es posible utilizar menos componentes qué los que se tienen, considerar un 0.85% de información o más.**
5. **Elaborar un modelo (modelo 1) de clasificación o regresión (logístico o máquina de vector soporte) considerando todas las variables que se proporcionan. Considere tener el mejor modelo posible (el qué tenga más accuracy).**
6. **Aplicar técnicas de selección de las variables o transformación de variables para decidir las variables más relevantes (PCA) del punto 5. En base al análisis del punto 5, realizar un nuevo modelo (modelo 2) de clasificación (logístico o máquina de vector soporte) con las variables seleccionadas, donde se puede suponer que la cantidad de variables es menor que el modelo anterior porque pasó por un proceso de selección de variables (Modelo 2).**
7. **Crear las conclusiones en base a los resultados de los 2 modelos. ¿Qué modelo es mejor? ¿Sirve reducir variables antes del modelado? ¿Se pudo o resulto útil reducir variables después del modelado?**

NOTA 1: La asignación de las bases de datos por equipo se encontrará en el archivo ‘Equipos_ProyectoFinal.xlsx’.

NOTA 2: Cada equipo debe de verificar la base de datos que le corresponde y si hubiera algún inconveniente (no se puede descargar, son muy pocos datos,

son muchos datos, la base de datos está dañada, etc.), se debe notificar con el profesor para realizar un cambio de la base de datos.

ENTREGABLES:

- **Reporte donde se explicará el proceso que se siguió para resolver el problema y se expondrán los resultados (en word o PDF).**
- **Código en jupyter notebook o en spyder**
- **Presentación, donde expondrán sus resultados**

Elementos básicos que debe de contener el reporte:

- El reporte del trabajo realizado se entregará de manera digital en un documento elaborado en computadora (word ó pdf, no fotografías de hojas ó cuadernos), e incluirá como mínimo:
  - ✓ Nombre y apellidos de alumno de los integrantes del equipo.
  - ✓ Es obligatorio que incluyan el código que generaron para la realización de la práctica (si adjuntan los archivos .py del código no es necesario que el código aparezca en su reporte; es suficiente que lo mencionen en alguna parte de su reporte y que indiquen el orden o forma hicieron la implementación de su algoritmo). La recomendación es que agreguen los códigos como apéndices en su reporte.
  - ✓ Es necesario que exista un capítulo que describa la solución que se propuso para el problema.
  - ✓ Por la asignación de datos a cada equipo, se espera que se obtengan resultados distintos en cada reporte (no copiar).
  - ✓ Incluyan figuras para visualizar los resultados.
  - ✓ Es obligatorio tener una sección de Introducción, desarrollo y conclusiones.
  - ✓ Se realizará una revisión de las faltas de ortografía.