**TITULO PROYECTO:**

**INTEGRANTES:**

1. **ENTENDIMIENTO DEL NEGOCIO**

**No dependen del científico de datos, por eso puede ser algo demorada esta etapa**

* 1. DESCRIPCIÓN DEL NEGOCIO: ¿a qué se dedica el negocio (especificidad)?
  2. DESCRIPCIÓN DEL PROBLEMA: un problema en específico que se quiera solucionar (eso lo da el experto)
  3. OBJETIVOS DE LA MINERÍA: (verbo en infinitivo) 🡪 Debe tener la tarea analítica específica, pueden tener varios, si se hará clustering, asociación y selección de factores, deben tener 3 objetivos
  4. DISEÑO DE SOLUCIÓN: tabla con el problema, tipo de análisis, tipo de aprendizaje, tarea analítica, requerimientos en los datos, métodos (específico, no poner etc 🡪 arma de doble filo), evaluación, Calidad esperada

**no se sigue hasta ser aprobado**

* 1. RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE  
     Qué se necesita en el despliegue? ¿Cómo será?  
     Servicio web? Nube? Jupyter Notebook?

Que pueda “conversar” el modelo con el despliegue

1. **ENTENDIMIENTO DE LOS DATOS**

**Datos específicos del problema (no es para todos los datos de la empresa), se hace en compañía con alguien del departamento de TI**

* 1. CICLO DE LOS DATOS  
     ¿Cómo se generan los datos?

¿Dónde se almacenan los datos?  
¿Quién puede modificar los datos? 🡪 mano humana es error, mientras más automatizado mejor  
¿Cada cuánto hay datos nuevos? Periodicidad 🡪 Permite hacer el cronograma de re entrenamiento

* 1. DICCIONARIO DE DATOS  
     Es una tabla con nombre de la variable, tipo y descripción de esta.  
     Esto es algo que se necesita por parte de la empresa, si no se tiene no se puede avanzar.
  2. REGLAS DE CALIDAD  
     No se construyen/calcula a partir de los datos.  
     Cada variable, ¿qué regla de calidad tiene asociada? Ejemplo: Edad 🡪 Clientes entre 18 y 60 años.  
     Limpieza de atípicos: No se puede hacer sin estas reglas de calidad.

LAS 2 PRIMERAS SON MUY COMPLICADAS YA QUE NO DEPENDEN DEL CIENTÍFICO DE DATOS

1. **PREPARACIÓN DE DATOS 🡪 Preparación estadística  
   Etapa técnica**
   1. INTEGRACIÓN  
      Crear una sábana de datos, hacer un JOIN, una sola tabla
   2. SELECCIÓN DE VARIABLES  
      Variables irrelevantes: No sirven para el análisis.  
      Protección de datos: Eliminar nombre, documento, dirección, teléfonos, etc.
   3. DESCRIPCIÓN ESTADÍSTICA  
      estadística descriptiva, media, moda, histogramas, caja de bigotes, diagramas de barras
   4. LIMPIEZA DE ATÍPICOS  
      Se hace con las reglas de calidad, valores por fuera del rango esperado (rango de calidad)
   5. LIMPIEZA DE NULOS  
      Imputación, no siempre se puede hacer. ¿Cuándo se puede o no hacer?
   6. CREACIÓN DE NUEVAS VARIABLES  
      Utilización de datos limpios
   7. ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA  
      Eliminar variables redundantes
   8. ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)  
      Cuándo una variable no me sirve. Ejemplo: Talla de zapatos para predecir la nota.
   9. REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)  
      Tiene riesgos, por eso es opcional.   
      Análisis de componentes principales (PCA).  
      Se hace para evitar maldición de la dimensionalidad.  
      Riesgo: Pérdida de información.  
      Una varianza muy diferente a la inicial (más del 25%) es una pérdida de los datos originales (no capturan la esencia original de estos datos).
   10. BALANCEO (CLASIFICACIÓN)  
       Se debe hacer con reglas especiales  
       Evaluación con datos ficticios 🡪 Desastre en el despliegue
   11. TRANSFORMACIONES  
       Skit-learn no maneja variables categóricas. Dependiendo del lenguaje se usan las reglas.
2. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN  
   No pueden ir aisladas (rendimiento).**
   1. CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING  
      Se aplican los métodos definidos en el diseño de la etapa 1.
   2. ANALISIS DE MEDIDAS DE CALIDAD  
      Análisis de las medidas de calidad (evaluación).
   3. SELECCIÓN DEL MEJOR MODELO

Complejidad computacional.  
Si la medida ROC de 2 métodos da 0.92 y 0.91, se elige según la complejidad computacional. ¿Diferencia computacional significativa?  
Uso de la ANOVA.

1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS   
      Despliegue con interfaz gráfica, aplicación móvil, nube, etc.
   2. CRONOGRAMA DE MANTENIMIENTO  
      Independientemente de que hayan o no errores, se debe de hacer mantenimiento del modelo en el tiempo.