**MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA**

**Entrega: Sábado 20 de junio**

**Límite: Sábado 13 de junio**

**Proyecto Integrador 1 (PI1) – 2019/2**

El PI1 corresponde al 35% de evaluación en cada una de las 3 materias: Almacenamiento y Recuperación de Información; Algebra de los datos; Estadística en analítica

**1. Estructura del entregable y fechas**

<propuesta> - documento

* <titulo>
* <integrantes,emails>
* <descripción-problema,contexto>
* <utilización-y-uso-de-una-metodología>: CRISP-DM o Microsoft-TDSP (colocar otras secciones de acuerdo a la metodología)
* <fuentes-datos>
* <entregables>
* <requerimientos-tecnológicos-o-definición-tecnología>
* <fechas-entregas>

1. COVID

Información Local > Nacional > Mundial

1. **Jara**
2. Covid Vs sentimientos
3. Covid vs Redes sociales
4. **David**
5. Covid Vs Estrategias de país
6. Covid Vs Economía (empleo, empresas)
7. **Valde**
8. Covid Vs Ayudas gobierno
9. Covid Vs Corrupción

**Catalina**

1. Covid Vs Sector salud
2. Covid Vs Acceso tecnológico

**Susana**

1. Covid Avance y mitigación (curva)
2. Covid Vs Transporte
3. Covid Vs Ambiente
4. Identificación de bodegas de Twitter: Uribista, Petrista, Fico.
5. Información de los clientes de banco y sugerencia de productos
6. Manejo del precio del petróleo
7. Corales

**Criterios de selección:**

Datos disponibles / Partiendo de la restricción

Alcance 40 horas máx/persona

Descripción

Baja complejidad de interpretación

<Avance1>

* definición del proyecto
* entendimiento del problema
* entendimiento de los datos
* entregables:
  + presentación pública
  + github del proyecto
  + reporte-técnico-y-modelos
* 2020/04/25

<Avance2>

* disposición tecnológica para el proyecto
* preparación de datos
* modelo(s) preliminar(res)
* entregables:
  + presentación pública
  + github del proyecto
  + reporte-técnico-y-modelos
* 2020/05/16

<Avance3>

* modelo(s) preliminar(res) validados
* entregables:
  + presentación pública
  + github del proyecto
  + reporte-técnico-y-modelos
  + producto-desplegado
* 2020/06/13

<PresentaciónyEvaluación>

* modelo(s) finales validados
* entregables:
  + presentación pública
  + github del proyecto
  + reporte-técnico-y-modelos
  + producto-desplegado

**2. Rúbricas generales de evaluación:**

2.1 Para la materia: ST1800 Almacenamiento y Recuperación de Información. 35%

* Ciclo de vida de los datos y procesamiento analítico (obligatorio)
  + Ambiente tecnológico: Servidores, Nube,  etc
  + Origen de los datos
  + Ingesta de datos (modos: batch, ETL, streaming)
  + Almacenamiento de los datos
    - Uso de sistemas de archivos distribuidos
    - Uso de bases de datos SQL o NoSQL de acuerdo a la problemática
  + Frameworks de procesamiento de datos.
  + Despliegue de modelos
    - Almacenamiento, motores SQL o NoSQL de acuerdo a la problemática
    - Persistencia de modelos y publicación en APIs.
* Si el problema está relacionado con procesamiento de texto, se evaluaría tambien:
  + Minería de texto
    - preparación de los datos (limpieza, tokenización, remoción de palabras de parada, stemming, lemmatization, POS.
    - modelos de representación de documentos, embedding, reducción de dimensionalidad
    - modelos de indexación, búsqueda, ranking, o diferentes técnicas en text mining como: tópicos, clasificación, agrupación, análisis de sentimientos, comunidades, etc, de acuerdo al caso.
    - validación y escalabilidad de los modelos

2.2 Para la materia: CM0888 ALGEBRA DE LOS DATOS. 35%

En las exposiciones y trabajos escritos del PI, los grupos deben presentar con detalle la implementación de los conceptos que se están tratando en la asignatura

* Limpieza y filtrado de datos a través de métricas y distancias adecuadas
* Identificación de outliers
* Aplicación de valores y vectores propios en la identificación de proyecciones y reducción de dimensión
* Normas y determinantes para calcular varianzas globales
* Productos internos para la identificación de asociaciones lineales entre variables.
* Aplicación de la descomposición en valores singulares
* Número condición de las matrices de covarianzas.

2.3 Para la materia: EC1801 ESTADÍSTICA EN ANALÍTICA. 35%

En las exposiciones y trabajos escritos del PI, los grupos deben presentar con detalle la implementación y aplicación de los conceptos que se están tratando en la asignatura

* Descripción de los datos multivariantes
* Regresiones y correlaciones
* Clasificación no supervisada y segmentación en grupos
* Clasificación supervisada con entrenamiento y validación