*Seminario de Proyectos*

*3B - Proyecto integrador - Objetivos específicos de investigación*

*Nombre: David Aaron Ramirez Olmeda*

*Programa: Maestría en Ciencia de Datos e Información*

**Introducción**

En esta unidad nos enfocaremos en definir los objetivos específicos para nuestro proyecto "Predicción anualizada de la tasa de accidentes viales en Sinaloa, México, mediante modelos avanzados de ciencia de datos". Antes de completar el ‘Constructo de investigación’ daremos primero una explicación a detalle de las ideas principales.

1. Desarrollar la base de datos y realizar el preprocesamiento:
   * Selecciona variables relevantes de los datos del INEGI sobre accidentes viales en Sinaloa, México, contenidas en nuestra basase de datos (archivos csv proporcionados por INEGI)
   * Limpiar los archivos csv, abordando inconsistencias evidentes, caracteres especiales y valores nulos para asegurar la integridad de la base de datos (usando Python).
   * Realizar unión de tablas pertinentes y discernir las relaciones entre nuestras ellas para enriquecer la información disponible para nuestro modelo.
   * Filtrar los datos para incluir únicamente información relacionada con el estado de Sinaloa.
2. Seleccionar y entrenar el algoritmo de predicción:
   * Evaluar distintos algoritmos de ciencia de datos, como XGBoost, Optimal Classification Trees (OCT), y modelos de ciencia de datos espacial.
   * Seleccionar el algoritmo más adecuado para la predicción de la tasa de accidentes viales, considerando su rendimiento en conjuntos de datos previamente preparados.
   * Ajustar el algoritmo seleccionado mediante técnicas de entrenamiento y validación cruzada (ajuste de hiperparámetros, análisi de overfitting, etc.).
   * Documentar y justificar la elección del algoritmo, proporcionando razones específicas para su preferencia (eficiencia, interpretación, etc.).
3. Analizar y contrastar resultados de los algoritmos:
   * Implementar el modelo ajustado en el conjunto de datos de prueba y evaluar su rendimiento.
   * Contrastar los resultados obtenidos utilizando diferentes algoritmos seleccionados anteriormente.
   * Realizar un análisis detallado de los valores numéricos generados por el modelo, incorporando métricas de evaluación de la predicción.
   * Explorar opciones visuales, como mapas, hotspots y estadísticas geolocalizadas, para enriquecer la interpretación de los resultados obtenidos.

Así, nuestro constructo de investigación queda de la siguiente manera:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Pregunta | Objetivo | Hipótesis |
| General |  |  | *Desarrollar un modelo de ciencia de datos que permita predecir la tasa de accidentes viales en Sinaloa, México.* |  |
| Específicos | 1 |  | Desarrollar la base de datos y realizar el preprocesamiento |  |
| 2 |  | Seleccionar y entrenar el algoritmo de predicción |  |
| 3 |  | Analizar y contrastar resultados de los algoritmos |  |

|  |  |
| --- | --- |
| Variable dependiente | Tasa de acciedentes viales en Sinaloa |
| Variables independientes | Género, edad, condiciones climáticas, tipo de carretera, posición geográfica, severidad, autoridades, aliento, heridas, muertes, etc. |