*Seminario de Proyectos*

*3C - Proyecto integrador - Pregunta general y específicas de la investigación*

*Nombre: David Aaron Ramirez Olmeda*

*Programa: Maestría en Ciencia de Datos e Información*

**Introducción**

En esta unidad nos enfocaremos en definir las preguntas específicas 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 y algunas ideas de nuestras preguntas.

**Pregunta General de Investigación:**

¿Cómo se puede desarrollar un modelo de ciencia de datos que permita predecir de manera efectiva la tasa de accidentes viales en el estado de Sinaloa, México, considerando variables relevantes y aplicando modelos avanzados de ciencia de datos?

**Preguntas Específicas de Investigación:** (En orden de relevancia)

1. Desarrollar la base de datos y realizar el preprocesamiento

* ¿Cuáles son los pasos esenciales para desarrollar la base de datos que sustentará el modelo de predicción de accidentes viales en Sinaloa, México?
* ¿Cómo se abordarán las inconsistencias, caracteres especiales y valores nulos presentes en el archivo CSV que contiene datos sobre accidentes?
* ¿En qué medida la unión de tablas adicionales enriquecerá la calidad de la base de datos y contribuirá a la precisión del modelo?

1. Seleccionar y entrenar el algoritmo de predicción

* ¿Qué criterios específicos justificarán la elección de algoritmos para la predicción de la tasa de accidentes viales en Sinaloa?
* ¿Cómo se aplicarán técnicas de entrenamiento, como la validación cruzada estratificada, para garantizar una representación equitativa de las clases en los conjuntos de entrenamiento y prueba?
* ¿Cuáles son los hiperparámetros clave que se ajustarán para optimizar el rendimiento del algoritmo seleccionado y mitigar el riesgo de sobreajuste?

1. Analizar y contrastar resultados de los algoritmos

* ¿Qué diferencias significativas se esperan al contrastar los resultados obtenidos utilizando diferentes algoritmos, considerando aspectos como interpretación, eficiencia y calidad predictiva?
* ¿Cómo se implementará el modelo ajustado en el conjunto de datos de prueba y cuáles serán las métricas utilizadas para evaluar su rendimiento?
* ¿Cómo se incorporarán elementos visuales, como mapas, hotspots y estadísticas en el análisis de los resultados para una comprensión más completa de la tasa de accidentes?

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Pregunta | Objetivo | Hipótesis |
| General |  | ¿Cómo se puede desarrollar un modelo de ciencia de datos que permita predecir de manera efectiva la tasa de accidentes viales en el estado de Sinaloa, México, considerando variables relevantes y aplicando modelos avanzados de ciencia de datos? | *Desarrollar un modelo de ciencia de datos que permita predecir la tasa de accidentes viales en Sinaloa, México.* |  |
| Específicos | 1 | ¿Cuáles son los pasos esenciales para desarrollar la base de datos que sustentará el modelo de predicción de accidentes viales en Sinaloa, México? | Desarrollar la base de datos y realizar el preprocesamiento |  |
| 2 | ¿Qué criterios específicos justificarán la elección de algoritmos para la predicción de la tasa de accidentes viales en Sinaloa? | Seleccionar y entrenar el algoritmo de predicción |  |
| 3 | ¿Qué diferencias significativas se esperan al contrastar los resultados obtenidos utilizando diferentes algoritmos, considerando aspectos como interpretación, eficiencia y calidad predictiva? | 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. |