**Escuela de Estudios de Postgrado**

**Facultad de Ingeniería**

**Universidad San Carlos de Guatemala**

**Maestría en Ingeniería para la Industria con**

**Especialización en Ciencias de la Computación.**

**Introducción a la Minería de Datos**

**M.A. Ingeniero Kevin Lajpop**

**PROYECTO**

**PARTE 1**

[**https://github.com/obduliogonzalez/proyecto\_1\_nogl**](https://github.com/obduliogonzalez/proyecto_1_nogl)

**Néstor Obdulio González López**

**Carné: 999013150**

**Proyecto 1 de Minería de Datos**

**Objetivo**

El objetivo de este proyecto es aplicar técnicas de minería de datos para analizar un conjunto de datos relacionado con faltas judiciales en Guatemala. A través de este análisis, se espera descubrir patrones y relaciones significativas, identificar grupos relevantes y desarrollar propuestas concretas que puedan contribuir a una mejor comprensión y abordaje de las problemáticas judiciales en el contexto guatemalteco.

**Conjunto de Datos**

El conjunto de datos utilizado en este proyecto corresponde a registros de faltas judiciales en Guatemala, disponible en el siguiente enlace: https://www.ine.gob.gt/bases-de-datos/faltas-judiciales/. Este dataset incluye información sobre los tipos de faltas, fechas, ubicaciones y otras variables relevantes para el análisis.

Fuente adicional de datos: Si es necesario, se buscarán fuentes adicionales para complementar el análisis y proporcionar una perspectiva más amplia sobre la problemática.

**Descripción General**

Este proyecto aplicará distintas técnicas de minería de datos para descubrir patrones y segmentar la información sobre faltas judiciales en Guatemala. Se abordarán los siguientes aspectos:

**Reglas de Asociación Apriori:** Descubrimiento de patrones frecuentes y relaciones entre variables.

**Reglas de Asociación FP-Growth:** Extracción de patrones adicionales y relaciones entre variables mediante otro enfoque de asociación.

**Análisis de Clúster (K-Means):** Segmentación de los datos en grupos relacionados para una mejor comprensión de las faltas judiciales y sus características.

**1. Reglas de Asociación Apriori**

Aplicando el algoritmo Apriori, se buscarán asociaciones entre variables del dataset.

**Configuración del algoritmo:** Definir el soporte y la confianza mínimos para identificar patrones significativos.

**Patrones interesantes identificados:** A continuación, se presentan algunos patrones descubiertos (se deben definir al menos 4):

**Patrón 1:**

"Las infracciones de tipo 1 (representadas por el código en Falta\_inf) tienden a ocurrir con mayor frecuencia en los meses de enero, octubre y noviembre, siendo el año 2013 el que presenta una frecuencia ligeramente mayor en comparación con el año 2014. Este patrón es particularmente notable en el grupo de edades de 20 a 39 años, quienes presentan el mayor número de registros. Además, las condiciones de alfabetismo y nacionalidad predominantes en estos casos incluyen los valores más comunes en el rango [1,9], lo cual sugiere un perfil de personas en situaciones de escolaridad básica."

Justificación del Patrón:

Frecuencia Temporal: Las gráficas de barras muestran que las infracciones registradas en Falta\_inf = 1 se concentran en los meses 1, 10 y 11, con el año 2013 presentando un número mayor de casos.

Distribución de Edad: El histograma de edad confirma que las infracciones de tipo 1 afectan mayormente a personas en los rangos de 20-29 y 30-39 años.

Condiciones Asociadas: La gráfica de red sugiere una fuerte asociación entre la infracción, los meses específicos y las variables de condición de alfabetismo (Cond\_alfabetismo\_inf) y nacionalidad (Nacionalidad\_inf), lo que podría estar relacionado con el contexto demográfico de los infractores.

Este patrón revela características y periodos específicos en los cuales se pueden diseñar intervenciones o medidas de control, considerando los perfiles demográficos y temporales observados en el análisis.

|  |
| --- |
|  |

**Patrón 2:**

Distribución por Departamento: La gráfica permite observar cómo los casos (representados como barras) se distribuyen en distintos departamentos a lo largo de los años (2013 y 2014). Algunos departamentos muestran una cantidad significativamente mayor de casos, lo que podría estar relacionado con factores específicos de cada región, como su tamaño poblacional o sus condiciones socioeconómicas.

Variación Anual: Se observa la tendencia de los casos en cada departamento a lo largo del tiempo. Por ejemplo, algunos departamentos podrían tener una cantidad constante de casos entre 2013 y 2014, mientras que en otros, los casos pueden haber aumentado o disminuido. Esta variación podría estar vinculada a políticas locales, cambios en la demografía, o intervenciones de salud y seguridad.

Diferencias por Sexo: Las barras de cada año están diferenciadas por sexo (hombre y mujer). En la mayoría de los departamentos, parece haber una distribución desigual entre sexos, donde uno de los géneros tiene mayor representación en los casos reportados. Esta disparidad puede señalar diferencias en el tipo de casos reportados, roles de género específicos en la sociedad, o diferencias en cómo hombres y mujeres interactúan con el sistema reportado en los datos.

Tendencias Regionales y Específicas: En ciertos departamentos, como los números 9 y 13, hay una alta concentración de casos en comparación con otros. Esto sugiere que estos departamentos pueden tener factores distintivos que resultan en un mayor número de casos. Estos factores podrían estar relacionados con la densidad poblacional, el nivel de urbanización o la actividad económica.

Posible Influencia de Factores Externos: Los datos sugieren que puede haber factores externos que afectan el número de casos por año y por sexo en cada departamento. Esto puede incluir cambios en políticas gubernamentales, acceso a servicios de salud y seguridad, y condiciones socioeconómicas particulares de cada región.

Interpretación de Implicaciones

Este patrón puede ser útil para identificar departamentos que requieren una mayor atención o recursos en políticas públicas y programas sociales específicos para cada sexo. Las diferencias entre años también pueden indicar el impacto de cambios recientes en políticas o eventos significativos en los departamentos afectados.

Esta visualización puede ser un punto de partida para investigaciones más profundas sobre los factores que influyen en la distribución de casos en los departamentos y la manera en que afecta a hombres y mujeres de manera diferencial.

|  |
| --- |
|  |
|  |

**Patrón 3:**

|  |
| --- |
|  |
|  |

**Patrón 4:** [Observación adicional]

**Interpretación y relevancia:** Explicar cómo estos patrones ayudan a comprender la problemática de las faltas judiciales y cómo podrían influir en decisiones o intervenciones. Por ejemplo, si se encuentra que ciertas faltas ocurren más en ciertas regiones o tiempos, podrían implementarse estrategias preventivas en esos lugares o períodos.

**2. Reglas de Asociación FP-Growth**

El algoritmo FP-Growth se aplicará para identificar más patrones de asociación, utilizando un método alternativo que puede descubrir patrones diferentes o adicionales a los hallados con Apriori.

**Configuración del algoritmo:** Establecer los parámetros de soporte y confianza.

Patrones interesantes identificados:

**Patrón 1:** [Describir la relación entre variables, por ejemplo, "Las faltas de tipo C suelen estar vinculadas con el factor Z."]

**Patrón 2:** [Otra relación significativa encontrada]

**Patrón 3:** [Observación importante]

**Patrón 4:** [Relación adicional relevante]

Interpretación y relevancia: Discutir cómo estos patrones complementan o refuerzan los hallazgos del Apriori y qué nuevas percepciones ofrecen sobre la problemática judicial.

**3. Análisis de Clúster**

Para el análisis de clúster, se empleará el algoritmo K-Means para identificar grupos dentro del conjunto de datos que compartan características comunes.

Configuración del algoritmo: Seleccionar el número de clústeres apropiado mediante el método del codo o similar.

**Grupos identificados:**

**Clúster 1:** [Describir las características de este grupo, como tipo de falta, frecuencia, ubicación.]

**Clúster 2:** [Características de este grupo]

**Clúster 3:** [Características adicionales]

Interpretación: Analizar cómo los grupos identificados pueden ayudar a entender y clasificar diferentes tipos de faltas judiciales. Por ejemplo, si se observa un clúster con faltas frecuentes en ciertas zonas, esto podría indicar la necesidad de reforzar las medidas de prevención en esas áreas.

**Visualización:** Incluir gráficos de los clústeres para una representación visual de los resultados.

**4. Propuestas**

A partir de los análisis de patrones y clústeres, se presentan propuestas para mejorar la situación de las faltas judiciales en Guatemala:

**Propuesta 1:** Implementación de campañas de prevención en las áreas con mayor frecuencia de faltas, según lo indicado por los patrones de asociación y clústeres.

**Propuesta 2:** Mejora en la seguridad en los períodos o áreas identificadas como de alto riesgo, basándose en los patrones temporales observados.

**Propuesta 3:** Desarrollo de programas de capacitación para las autoridades locales en las zonas con mayor incidencia de faltas, promoviendo estrategias de intervención temprana.

**Justificación y viabilidad:** Cada propuesta se basa en los hallazgos del análisis de datos y se valida con documentación científica y un análisis del contexto guatemalteco, para asegurar su aplicabilidad y relevancia.