**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:**

Distribución de Niveles de Escolaridad:

Las gráficas muestran que ciertos niveles de escolaridad tienen mayor presencia en algunos departamentos y en ciertos años. Por ejemplo, en algunos departamentos es evidente que un nivel específico predomina a lo largo de los años, lo que sugiere una tendencia de permanencia o concentración de individuos con ese nivel educativo.

En los gráficos mejorados, al resaltar el nivel de escolaridad más frecuente, se facilita la observación de las áreas donde ciertos niveles son dominantes.

Tendencias Temporales:

La inclusión de los años permite identificar patrones de cambio o estabilidad en los niveles de escolaridad. Por ejemplo, si en un departamento el nivel de escolaridad dominante cambia de un año a otro, podría sugerir una variación en las políticas de contratación, migración de personas con distinto nivel educativo, o cambios en las oportunidades de educación en la región.

Si los niveles de escolaridad se mantienen constantes en todos los años, indica una estabilidad en el perfil educativo de la población en esos departamentos.

Comparación entre Sexo y Escolaridad:

Las gráficas específicas de sexo permiten ver si hay una diferencia notable en la distribución de niveles de escolaridad entre hombres y mujeres. Esto ayuda a identificar si existe una tendencia en la escolaridad que esté vinculada al género, lo cual puede ser relevante para políticas de inclusión o capacitación enfocadas en cada sexo.

Interpretación del Patrón

Este patrón es útil para identificar:

Áreas de Intervención: Departamentos que tienen un nivel de escolaridad bajo predominante pueden beneficiarse de programas educativos o de capacitación laboral.

Tendencias por Año: Los cambios en el nivel de escolaridad predominante pueden indicar variaciones en la educación o migración de personas, lo que podría afectar la planificación de recursos en diferentes departamentos.

Diferencias de Escolaridad por Sexo: Si se encuentran disparidades significativas en la escolaridad entre hombres y mujeres, esto podría ser indicativo de barreras de acceso a la educación que afectan de manera desproporcionada a un sexo.

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

**Patrón 4:**

Este patrón permite analizar la distribución de casos entre hombres y mujeres en distintos departamentos y a lo largo de los años. Es útil para identificar:

Diferencias por Departamento: Cuántos casos hay en cada departamento y si hay una predominancia por sexo.

Variaciones Anuales: Cambios en la cantidad de casos año a año en cada departamento, permitiendo ver tendencias o anomalías.

Comparación por Sexo: Visualización de posibles disparidades entre hombres y mujeres en distintos departamentos y años.

Este tipo de gráfico facilita el análisis del contexto geográfico (departamento), temporal (año) y demográfico (sexo) de los casos.

|  |
| --- |
|  |
|  |

**Interpretación y relevancia:** La combinación de estos patrones permite construir una visión integral de la problemática de faltas judiciales, considerando factores como el tipo de falta, ubicación geográfica, nivel educativo, profesión, sexo y cambio en el tiempo. Estos análisis pueden influir en decisiones estratégicas al:

Identificar Áreas de Mayor Incidencia: La localización geográfica de las faltas y su variación en el tiempo ayudan a los responsables de políticas a focalizar esfuerzos en áreas específicas.

Entender Factores de Riesgo Demográficos y Sociales: Patrones como la relación entre profesión, nivel educativo y sexo con las faltas permiten identificar grupos de riesgo, lo que facilita el diseño de intervenciones específicas.

Evaluar la Efectividad de Intervenciones Actuales: Los análisis temporales permiten observar si las incidencias de faltas están disminuyendo o aumentando tras intervenciones previas, lo cual es clave para ajustar o rediseñar estrategias.

Diseñar Programas Preventivos y Educativos: La comprensión de los factores asociados a las faltas judiciales puede fundamentar programas educativos que aborden tanto la prevención de las faltas como el fomento de comportamientos pro-sociales.

**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:**

1. Selección del Número Óptimo de Clústeres (Método de la Silueta)

La gráfica del método de la silueta muestra que el valor de 𝑘 = 2 tiene el mayor ancho promedio de la silueta, lo que indica que los clústeres están mejor definidos cuando se dividen en dos grupos. Sin embargo, también observamos una estructura aceptable de clústeres para k=4, aunque con un valor de silueta menor.

La decisión de utilizar k=4 en lugar de k=2 puede deberse a la necesidad de capturar más subgrupos en los datos, aunque con una menor cohesión dentro de cada clúster. Esto es algo común en análisis exploratorios cuando se busca segmentar los datos en más categorías para obtener una mayor comprensión de las diferencias entre subgrupos.

2. Visualización de Clústeres con K-Means (k=4)

La segunda gráfica muestra los datos proyectados en dos dimensiones (Dim1 y Dim2) usando una técnica de reducción de dimensionalidad, como PCA, para visualizar los clústeres obtenidos con 𝑘 = 4.

Clúster 1 (color rojo):

Este grupo está ubicado en la parte inferior izquierda de la gráfica.

Puede representar un grupo con características que difieren en gran medida de los otros, ya que se encuentra separado en la dimensión 1.

Clúster 2 (color verde):

Ubicado en la parte central, este clúster tiene una posición intermedia y parece tener una ligera superposición con otros grupos.

Podría representar un grupo con características promedio o intermedias en comparación con los otros clústeres.

Clúster 3 (color azul claro):

Este grupo se extiende a lo largo del eje positivo en Dim1 y Dim2, ocupando una porción considerable de la gráfica.

La extensión de este clúster sugiere que podría ser un grupo con más variabilidad interna o que contiene muchos puntos con características comunes.

Clúster 4 (color lila):

Este grupo está en la parte superior izquierda, en una zona distinta a los otros clústeres, lo cual indica características particulares.

La separación de este clúster indica que puede representar un subgrupo específico con características únicas en comparación con el resto.

Resumen de los Clústeres y Sugerencias

Interpretación de Clústeres: Cada clúster representa un subgrupo dentro de los datos que comparte características similares en las dimensiones seleccionadas (en este caso, una combinación de variables que fueron escaladas y proyectadas en las dimensiones principales).

Cohesión de los Clústeres: Dado el valor de la silueta, es probable que los clústeres no sean perfectamente cohesivos con k=4, lo que indica que podría haber superposición entre algunos clústeres. Esto es común en datos con patrones complejos.

Posibles Acciones: Puedes realizar una revisión adicional de los centroides de cada clúster para describir mejor las características de cada grupo y confirmar si los clústeres resultantes tienen sentido práctico.

|  |
| --- |
|  |
|  |

**Clúster 2:**

Observaciones Generales de la Visualización

Distribución en el Plano Dim1 y Dim2: La gráfica muestra cuatro clústeres claramente diferenciados en el plano proyectado (Dim1 y Dim2). Cada elipse representa la dispersión de puntos dentro de cada clúster en función de las dos dimensiones principales, que explican un 43.3% y 32.3% de la varianza, respectivamente.

Tamaño y Forma de los Clústeres:

El tamaño de las elipses indica la variabilidad interna dentro de cada clúster.

Los clústeres que tienen una elipse más grande pueden tener una mayor variabilidad en sus características internas.

Los clústeres que tienen elipses más alargadas, como el Clúster 2 (en verde), pueden estar extendidos a lo largo de una dimensión, lo que sugiere que los datos en ese grupo tienen más diferencias en una de las características.

Análisis del Clúster 2 (Verde)

Posición y Forma:

El Clúster 2 se encuentra en el centro de la gráfica, lo que indica que, en términos de Dim1 y Dim2, podría representar una "posición intermedia" entre los demás clústeres. Esto sugiere que los datos en este clúster pueden tener características que son un punto medio entre los otros grupos.

La forma alargada del elipse verde sugiere una dispersión de datos a lo largo de una de las dimensiones, lo cual podría indicar una diversidad en una característica específica dentro de este clúster.

Tamaño de la Elipse:

El Clúster 2 tiene una elipse relativamente angosta, lo que podría indicar que los datos dentro de este grupo están relativamente concentrados a lo largo de una dimensión y tienen menor variabilidad en la otra.

Esto puede sugerir que, aunque hay diversidad en alguna característica específica, el grupo es relativamente homogéneo en otros aspectos.

Comparación con Otros Clústeres:

Comparado con el Clúster 3 (en azul), que tiene una elipse grande y extendida, el Clúster 2 parece ser más concentrado, lo que sugiere una menor variabilidad interna.

La posición del Clúster 2 entre el Clúster 1 (rojo) y Clúster 3 (azul) indica que comparte ciertas similitudes con estos grupos, pero tiene un perfil distinto.

Posible Interpretación del Clúster 2

Basado en la posición y forma del Clúster 2, podríamos inferir lo siguiente:

Característica Media/Intermedia: El Clúster 2 podría representar individuos o elementos con características promedio o intermedias en comparación con los otros clústeres.

Diversidad en una Dimensión: La forma alargada de la elipse sugiere que hay variabilidad en una característica particular dentro de este grupo. Esto podría representar diferencias en alguna característica específica (por ejemplo, rango de edad o nivel de falta) que es diversa dentro del grupo.

Posible Grupo de Transición: Dado que este clúster se encuentra en una posición intermedia en la gráfica, podría representar un grupo de transición, es decir, individuos o elementos que comparten ciertas características con otros grupos pero no encajan completamente en ellos.

Interpretación de los Resultados para Clúster 2

Característica Clúster 1 Clúster 2 Clúster 3 Clúster 4

Edad\_inf -0.316 -0.306 3.200 -0.315

Depto\_boleta -0.810 0.069 -0.610 0.864

Falta\_inf 0.774 -1.044 -0.659 0.642

Dado que los valores están escalados (estandarizados), aquí tienes la interpretación basada en las diferencias relativas entre los clústeres:

Edad\_inf (Clúster 2: -0.306):

El valor estandarizado de Edad\_inf en el Clúster 2 es negativo y cercano a cero, lo cual indica que los individuos de este grupo tienen una edad levemente por debajo del promedio de todo el conjunto de datos, pero no significativamente más jóvenes o mayores.

Comparado con el Clúster 3 (que tiene un valor alto de 3.200), el Clúster 2 representa a personas de edad promedio en el contexto del dataset.

Depto\_boleta (Clúster 2: 0.069):

El valor cercano a cero para Depto\_boleta en el Clúster 2 indica que las personas en este grupo no están concentradas en departamentos específicos. Esto sugiere que el Clúster 2 podría tener una dispersión geográfica mayor o incluir individuos de áreas promedio en el contexto del dataset.

A diferencia del Clúster 4, que tiene un valor alto en Depto\_boleta (0.864), y el Clúster 1, con un valor negativo significativo (-0.810), el Clúster 2 parece representar una mezcla geográfica.

Falta\_inf (Clúster 2: -1.044):

El valor negativo en Falta\_inf sugiere que los individuos en el Clúster 2 tienen un número de faltas menor que el promedio general del conjunto de datos.

Este clúster podría representar a personas con un historial de infracciones bajo, en contraste con el Clúster 1 (0.774) y Clúster 4 (0.642), que tienen valores positivos y sugieren más faltas.

Resumen del Perfil del Clúster 2

Edad: Personas de edad promedio o ligeramente por debajo del promedio.

Ubicación Geográfica: Distribución geográfica diversa o representación promedio.

Nivel de Faltas: Menor cantidad de faltas en comparación con los otros clústeres.

Interpretación General del Clúster 2

El Clúster 2 parece representar un grupo de personas con características promedio en cuanto a edad y ubicación geográfica, pero con un historial de faltas significativamente menor que otros grupos. Esto podría ser útil si estás buscando un subgrupo con bajo nivel de infracciones o buen comportamiento en términos de faltas.

|  |
| --- |
|  |

**Clúster 3:**

1. Edad\_inf (Media: 3.20, Desviación Estándar: 0.00)

La media de 3.20 para Edad\_inf en Clúster 3 es significativamente mayor en comparación con los otros clústeres.

La desviación estándar de 0.00 indica que todas las observaciones en el Clúster 3 tienen el mismo valor escalado de Edad\_inf. Esto significa que las edades en este clúster son homogéneas y consistentemente altas, en comparación con los otros grupos.

Dado que el valor está escalado, en términos prácticos, podríamos interpretar que este clúster representa a un grupo de personas de edad considerablemente mayor en comparación con los demás clústeres.

2. Depto\_boleta (Media: -0.61, Desviación Estándar: 0.84)

La media de -0.61 en Depto\_boleta sugiere que las personas en este clúster tienden a estar ubicadas en departamentos que están representados por valores más bajos en el contexto del dataset.

La desviación estándar de 0.84 indica una variabilidad moderada, lo que significa que aunque este grupo tiende a estar en departamentos con valores negativos (departamentos específicos), no todos los individuos del clúster están en un solo lugar. Hay cierta dispersión dentro de este grupo.

3. Falta\_inf (Media: -0.66, Desviación Estándar: 0.94)

La media de -0.66 en Falta\_inf sugiere que las personas en el Clúster 3 tienden a tener un nivel de faltas por debajo del promedio general. Este clúster podría estar compuesto por personas que cometen menos infracciones en comparación con otros grupos.

La desviación estándar de 0.94 indica una mayor variabilidad en Falta\_inf, lo cual sugiere que, aunque el grupo tiende a tener menos faltas en general, hay una diversidad de comportamientos en términos de infracciones dentro del clúster.

Resumen del Perfil del Clúster 3

Con base en los análisis anteriores, el Clúster 3 se puede interpretar de la siguiente manera:

Grupo de Edad Avanzada: Este clúster representa personas de una edad significativamente mayor en comparación con los demás clústeres, con una edad muy homogénea.

Ubicación en Departamentos Específicos: Los individuos de este clúster están en departamentos con valores bajos en el dataset, pero existe una dispersión moderada, indicando que no todos están en un único departamento.

Nivel Bajo de Infracciones: Tienen un nivel de faltas por debajo del promedio, aunque con cierta variabilidad. Esto sugiere que, en general, este es un grupo con buen comportamiento o pocas infracciones, aunque algunos pueden tener un historial ligeramente diferente dentro del clúster.

Conclusiones

Este clúster podría representar a un grupo de personas de mayor edad, posiblemente con más experiencia o estabilidad, lo cual puede estar relacionado con un menor número de infracciones. Este perfil podría ser relevante si se están buscando segmentos con bajo riesgo de faltas o con características de comportamiento estables.

|  |
| --- |
|  |
|  |
|  |

**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.