Daniel Esteban Aguilera Figueroa – 202010592
Laura Vanessa Martínez Prieto – 202012624
Cristian Armando Sánchez Ocampo – 202022112

"Informe: Laboratorio 1"

1. Introducción

Para este laboratorio se generará una caracterización e identificación de posibles patrones que pueden ser utilizados para entender los factores que indicen a la severidad de los accidentes con el objetivo de trazar mapas de seguridad vial. Se espera aplicar tres diferentes algoritmos de clustering para lograr resolver el objetivo entregado por la organización, se generarán conclusiones a partir de los modelos construidos y se comunicaran los análisis explicando el valor que podrían tener para la organización.

2. Entendimiento de los datos

En esta parte del análisis se realizó el entendimiento total de los datos entregados, para los cuales se nos fue entregado un CSV que contenía la información de los accidentes recopilados por BiciAlpes, en conjunto fue agregado un diccionario el cual contenía la información de las variables tenidas en cuenta en cada accidente. Este momento fue dedicado en su totalidad a entender las 14 variables presentadas y 5338 entradas. De estas, tres son numéricas:

- Number_of_Vehicles, Number_of_Casualties, Speed_limit

Adicionalmente, se tienen 11 variables categóricas:

- Accident_Severity, Day_of_Week, Time, Road_Type, Junction_Detail, Light_Conditions, Weather_Conditions, Road_Surface_Conditions, Urban_Or_Rural_Area, Vehicle_Type, Did_Police_Officer_Attend_Scene_Of_Accident.

La información relacionada a estas puede ser encontrada en la carpeta adjunta "datos" en el archivo "Diccionario_BiciAlpes.xlsx" proporcionado.

3. Preparación de los datos

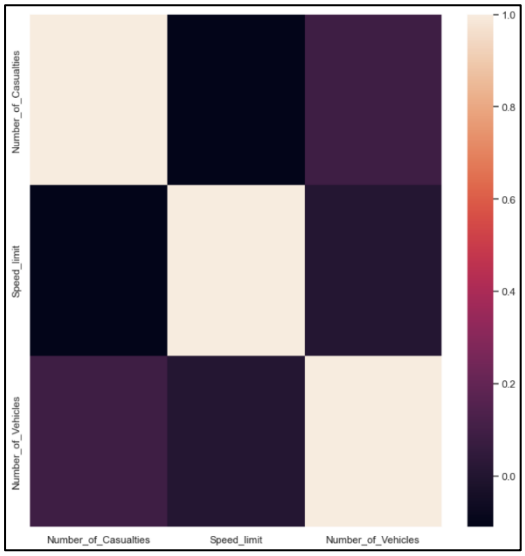
Para esta fase del laboratorio teniendo en cuenta las dimensiones de calidad de datos en los que se toman en cuenta cuatro ítems importantes:

- Completitud: Para esta primera dimensión se tuvo en cuenta la cantidad de valores nulos presentes en el dataset y el contexto del negocio, se generó una limpieza en estos datos en el cual fueron afectadas dos columnas: "Day_of_Week" la cual presentaba 0.0035 datos nulos y "Unnamed: 14" la cual se decidió eliminar por completo ya que el 100% de sus datos no tenían ningún valor relevante.
- Unicidad: Para esta parte se decidió el no realizar ninguna alteración en los datos, ya que teniendo en cuenta el contexto del negocio y los datos entregados no es posible identificar si existe algún tipo de dato duplicado o si existiera no afectaría en ningún aspecto el análisis de los resultados esto se debe a que los accidentes no cuentan con un identificador único.
- Consistencia: En cuanto a la consistencia no se presenta algún tipo de problema que afecte considerablemente la consistencia en el dataset ya que una gran parte de las variables de tipo categórico se representan con diferentes números que

representan valores distintos. En caso de aquellas que estaban representadas con Strings, se procedió a hacer una numeración “label encoding” para generar consistencia.

- Validez: Para esta parte se decidió hacer una comprobación con la función unique() y así lograr comprobar que efectivamente los datos originales entregados cumplen con esta dimensión.

4. Modelamiento



Para esta fase del proceso se comenzó generando un mapa de calor sobre las variables de tipo numérico para así determinar la correlación entre cada una de ellas, el resultado fue el siguiente:

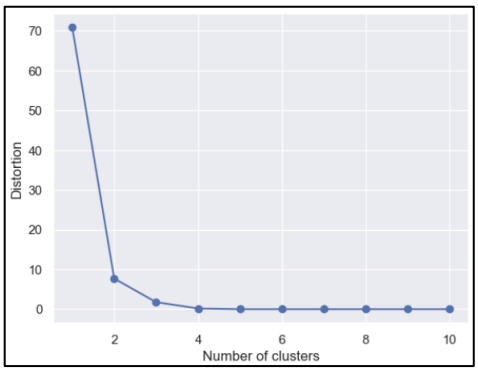
Partiendo de este diagrama se logra analizar un poco más a fondo el comportamiento de los datos, por lo que se decidió modelar los datos partiendo de implementación de los algoritmos k-means, k-modes y Aglomerative Clustering. A continuación, se presentará una pequeña descripción de estos, se realizará su respectivo análisis y comparación.

4.1. Algoritmos

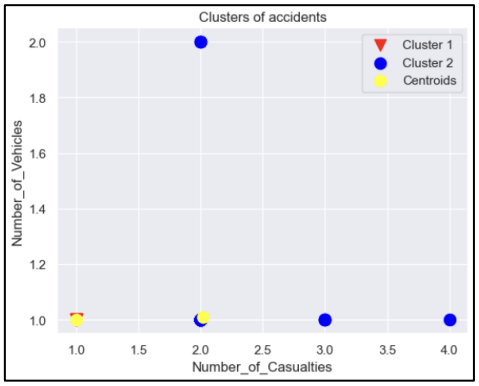
- K-means – Cristian Sánchez

Este algoritmo iterativo se basa en dividir cualquier grupo de datos en k clústers distintos en donde cada uno de los datos solo puede pertenecer exclusivamente a un grupo. Este algoritmo en específico fue usado para analizar el comportamiento de las variables numéricas, se usó el método del codo para identificar la cantidad de clústeres a generar. Se decidió generar un gráfico distinto para cada una de las posibles combinaciones entre las variables y este fue el resultado:

- Number_of_Vehicles vs Number_of_Casualties

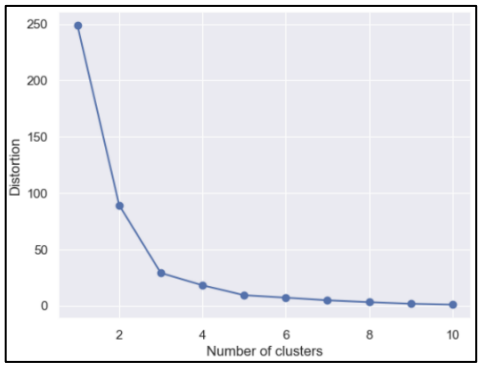


Para este caso es bastante evidente que el número de clusters a generar es 2

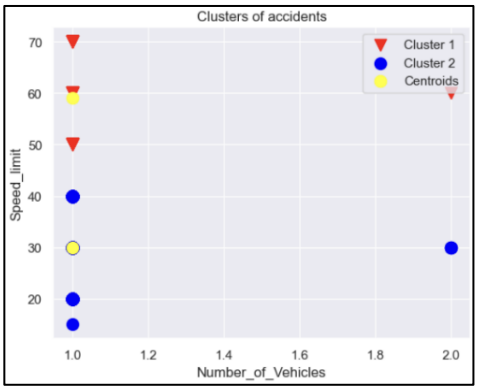


Para este caso se puede observar la diferencia entre la varianza de los dos clusters obtenidos. Para el cluster1 es notable que la varianza de muy mínima ya que los datos se encuentran en un mismo punto, a diferencia del cluster2 el cual tiene una varianza más notable ya que los datos pertenecientes a esta están a una distancia considerable.

- Speed_limit vs Number_of_Casualties



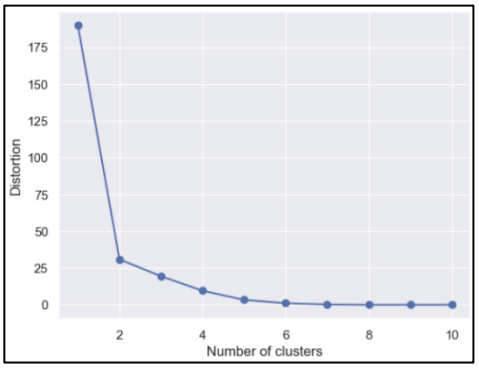
Para este caso es bastante evidente que el número de clusters a generar es 2



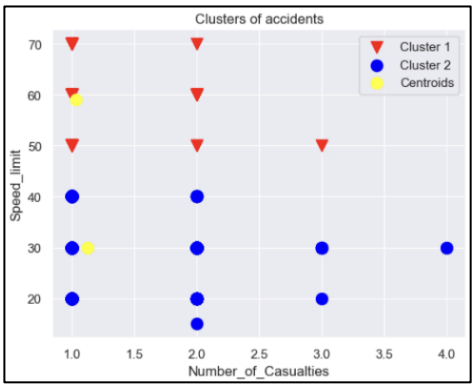
Para este segundo escenario se puede observar que los datos varían considerablemente en el eje de la variable “Speed_limit”, pero así

mismo que los datos en el eje de “Number_of_Casualities” están tendiendo a uno.

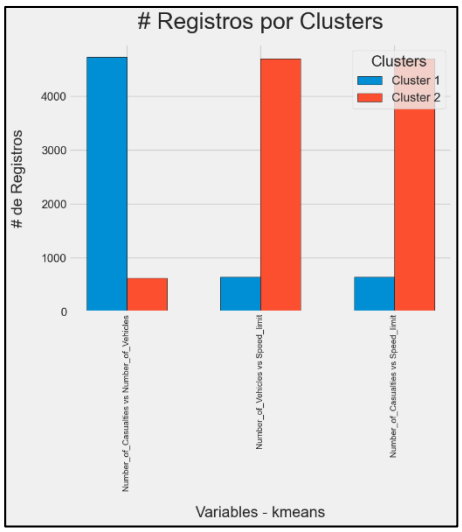
- Speed_limit vs Number_of_Vehicles



Para este caso es bastante evidente que el número de clusters a generar es 2

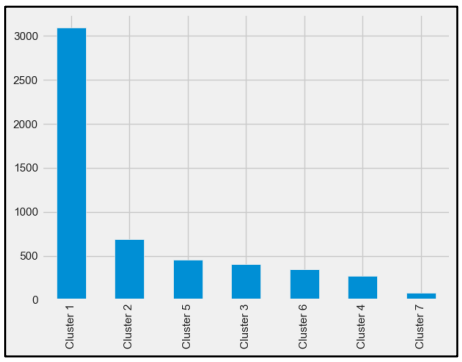


Para este último caso también se puede un evidenciar un comportamiento variado en el eje correspondiente a “Speed_limit”, pero con una mayor tendencia a uno en el eje de “Number_of_Casualities”.

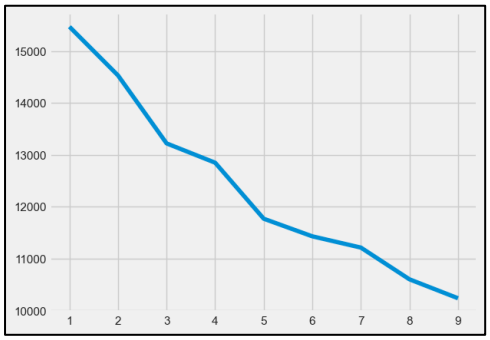


- K-modes – Daniel Aguilera

    - Es un método de aprendizaje no supervisado el cual se basa en dividir un conjunto de datos en diferentes grupos, de tal forma que cada uno de los datos pertenecientes a un grupo sean lo más similares posible. Este método es usado generalmente en variables categóricas de forma que su clustering corresponda a un resultado esperado como el mostrado a continuación.

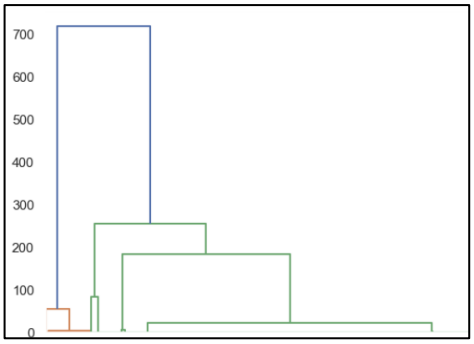


Es importante que para llegar a este número de clusters se realizaron distintas pruebas usando el método del codo.
Aunque tiene distintas particiones, se tuvieron mejores resultados con siete clusters.

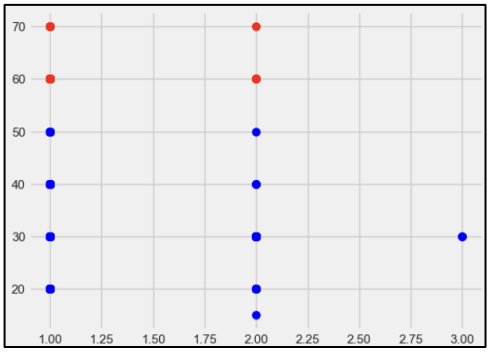


- Aglomerative Clustering – Laura Martínez

    - Este algoritmo se encarga de agrupar datos con diferentes cualidades similares en grupos, para este enfoque en específico se comienza desde clústeres más pequeños hasta crear clústeres más grandes. Se decidió implementar este algoritmo en el caso de Speed_lmit vs Number_of_Casualties, para así lograr un mejor análisis y lograr una comparación con el grafico entregado por el algoritmo de k-means.
    Para definir el número de clusters a realizar se decidió generar un dendograma, el cual se puede observar a continuación:



Es importante tener en cuenta que para este paso se vio necesario generar los gráficos con un segmento de los datos ya que las maquinas generan error al correr una cantidad de datos mayor a tres mil.
Teniendo en cuenta el dendograma es notable que también se generan dos cluster, y es posible realizar la siguiente representación:

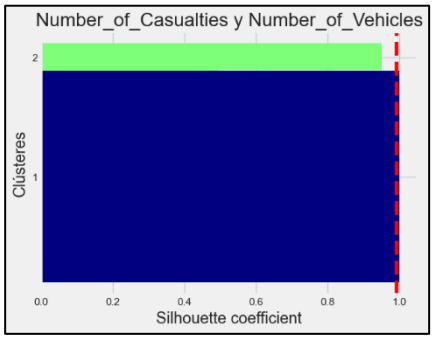


De igual forma que el generado por el algoritmo de k-means, este algoritmo entrega un gráfico en donde los datos son variados con respecto al eje de “Speed_limit” y datos que tienden a uno en el eje de “Number_of_Casualties”.

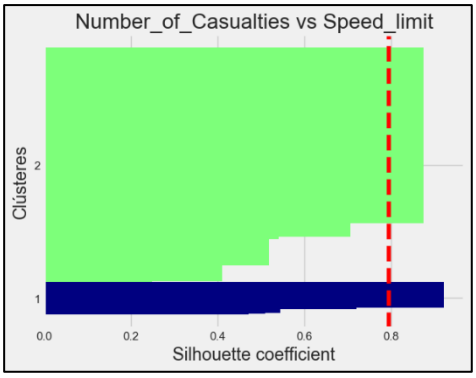
5. Validación

5.1. Cuantitativa

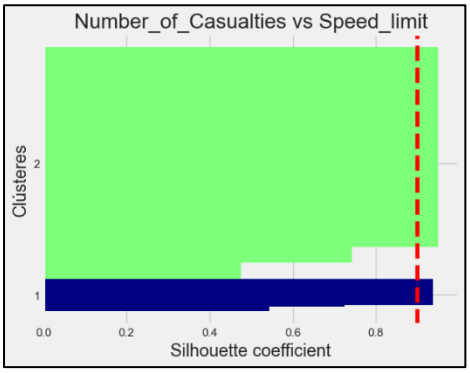
- Para la validación de los algoritmos de forma cuantitativa se decidió usar, por un lado, el método del codo, el cual fue evidenciado en los análisis realizados anteriormente. Adicionalmente, usando el coeficiente de silueta obtuvo un análisis a profundidad, encontrando de alguna forma, que tan bueno fue el modelo.
- K-means
  a. Number_of_Vehicles vs Number_of_Casualties



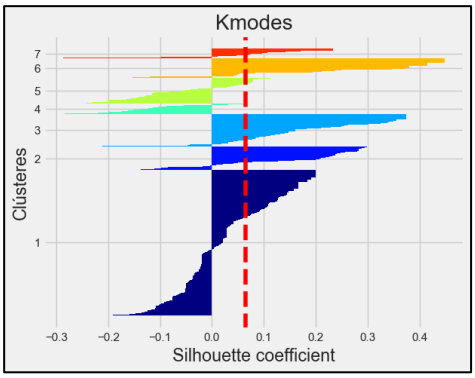
b. Speed_limit vs Number_of_Casualties



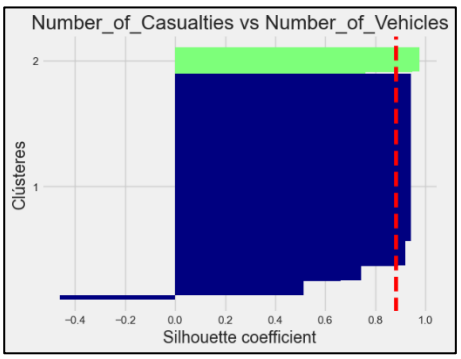
c. Speed_limit vs Number_of_Vehicles



- Kmodes



- Agglomerative



- Al analizar los algoritmos implementados se identifica una distribución mejor agrupada por parte de K-means y el aglomerativo dadas las siluetas cercanas a 1

(Aunque Aglomerative muestra una pequeña desviación en el primer cluster). Sin embargo, k-means se implementa con los datos numéricos y el aglomerativo con un conjunto más reducido de datos por la complejidad de este algoritmo. Por otro lado, el algoritmo K-modes permite un mejor análisis sobre los datos, ya que se tiene en cuenta las variables categóricas (factores que influyen en la accidentalidad) y la totalidad de los datos y, aunque no se haya tenido una buena distribución en los clusters de este, permitió sacar conclusiones más significativas sobre los factores que influyen en los accidentes.

5.2. Cualitativa

Por lo anterior, se puede saber que la mayoría de los accidentes son de bicicletas y que por lo general son leves, además que los principales factores presentes en la mayoría de los accidentes son el tipo de carretera (rotondas y sus derivadas), días laborales y las condiciones climáticas (presencia de lluvia o vientos leves). Adicionalmente, aunque la agrupación de los accidentes no haya sido la esperada, los datos importantes corresponden a un análisis del siguiente estilo, teniendo en cuenta la gravedad de los accidentes.

- Para Junction Detail, se tiene una gran influencia en la severidad de los accidentes cuando se presenta en una rotonda o minirrotonda
- Para Weather Conditions, se tiene una gran influencia en la severidad de los accidentes cuando se presentan vientos suaves
- Para Road Surface Conditions, se tiene una gran influencia en la severidad de los accidentes cuando se presentan condiciones secas
- Para Light Conditions, se tiene una gran influencia en la severidad de los accidentes cuando se presentan condiciones de luz diurna
- Para Urban_or_Rural_Area, se tiene una gran influencia en la severidad de los accidentes cuando se presentan en zonas urbanas
- Para Did_Police_Officer_Attend_Scene_of_Accident, se tiene una gran influencia en la severidad de los accidentes cuando se presentan en accidentes donde si se presentó un oficial de policía
- Para Day_of_Week, se tiene una gran influencia en la severidad de los accidentes cuando se presentan en los días laborales
- Para Time, se tiene una gran influencia en la severidad de los accidentes cuando se presentan en la tarde

6. Bibliografía

- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/
- https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019
- https://harikabonthu96.medium.com/kmodes-clustering-2286a9bfdcfb

7. Observaciones

El documento del informe se encontrará en la carpeta “docs” al igual que los demás entregables importantes.