UNIVERSIDAD ANDINA DEL CUSCO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

DETECCIÓN DE PATRONES DE PERSONAS DESAPARECIDAS MEDIANTE TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

Presentado por: Rimachi Costillas, Roy Joseph

Para optar el Título Profesional de Ingeniero de Sistemas

Asesor: Ing. Iván Molero Delgado

DEDICATORIA

A mi madre y a mi hermano, por ser la fuente principal de mi motivación que me ayuda a mejorar como persona y profesional cada día.

AGRADECIMIENTOS

Agradezco a mi familia, por apoyarme y darme su soporte en todo mi desarrollo como profesional.

También agradezco a mi asesor Mgt. Ivan Molero Delgado por apoyarme en todo el proceso investigativo y a resolver las dudas, a la Mgt. Pilar Hidalgo Leon por guiarme y motivarme en el campo de la investigación, además de compartir su experiencia en el campo que me ayudo a concretar la investigación.

RESUMEN

La desaparición de personas es una de las preocupaciones principales tanto a nivel nacional como mundial, estás se pueden dar debido a la trata de personas, tráfico de órganos, entre otros. Dentro de los grupos de personas desaparecidas existe uno cuyas características alertan más a la sociedad, por lo cual requieren una respuesta más rápida y eficiente; a este grupo se le denomina personas en situación de vulnerabilidad y está conformado por niños, niñas, adolescentes, personas adultas mayores y personas con discapacidad física, mental o sensorial.

El aprendizaje no supervisado por otro lado forma parte del aprendizaje automático que a su vez es parte del campo de la Inteligencia Artificial, esta rama busca recolectar o generar conocimiento a través de la información albergada en los datos sin la necesidad de etiquetarlos. Los algoritmos de aprendizaje no supervisado cotidianamente son parte de soluciones tecnológicas que permiten segmentar o descubrir patrones de un conjunto de datos. Dichos patrones han servido a múltiples campos para desarrollar estrategias focalizadas por grupo, incrementando así la eficacia de los procesos que se encargan de combatir una problemática determinada.

Los datos recolectados de menores desaparecidos contienen múltiples atributos como: edad, genero, raza, color de ojos, color de cabello, tipo de nariz, tipo de boca, etc. Entre estos campos solo existe una etiqueta cuyo valor puede ser “desaparecido” o “encontrado”, esta etiqueta no solventa la aplicación de técnicas de aprendizaje supervisado; debido a esto se opto por utilizar técnicas de aprendizaje no supervisado que surgen como una alternativa viable para analizar los datos. Además, este tipo de aprendizaje debido a su enfoque que no requiere de etiquetas en los datos disminuye el costo de recursos. Por esta razón la investigación busca describir o mostrar conocimiento sobre los patrones que puedan ser detectados dentro del conjunto de datos haciendo uso de las técnicas de aprendizaje no supervisado.

Por consiguiente, para aplicar las técnicas de aprendizaje no supervisado primero fue necesario extraer todos los datos albergados en la página web utilizando la técnica de web scraping que nos permitió obtener todos los datos sobre el perfil del menor. También, debido a que el conjunto de datos recolectado contenía inconsistencias entre sus registros, se preprocesaron con técnicas del proceso KDD para obtener la mayor cantidad de registros validos para el estudio.

Finalmente, el análisis de los datos se llevo a cabo variando entre múltiples números de clústeres determinados por el método del codo, para así pasarlos al algoritmo k-means y así determinar mediante métricas de validación la cantidad adecuada para el conjunto de datos.

ABSTRACT

The disappearance of people is one of the main concerns both nationally and globally, these can occur due to human trafficking, organ trafficking, among others. Within the groups of disappeared persons there is one whose characteristics alert society more, for which they require a faster and more efficient response; This group is called people in vulnerable situations and is made up of boys, girls, adolescents, older adults and people with physical, mental or sensory disabilities.

Unsupervised learning on the other hand is part of machine learning which in turn is part of the field of Artificial Intelligence, this branch seeks to collect or generate knowledge through the information stored in the data without the need to label it. Unsupervised learning algorithms daily are part of technological solutions that allow you to segment or discover patterns in a data set. These patterns have served multiple fields to develop group strategies, thus increasing the effectiveness of the processes that are responsible for combating a specific problem.

The data collected from missing minors contains multiple attributes such as: age, sex, race, eye color, hair color, type of nose, type of mouth, etc. Among these fields there is only one label whose value can "disappear" or "found". This label does not address the application of supervised learning techniques; Due to this, it was decided to use unsupervised learning techniques that emerge as a viable alternative to analyze the data. In addition, this type of learning due to its approach that does not require labels on the data reduces the cost of resources. For this reason, the research seeks to describe or show knowledge about the patterns that can be detected within the data set using unsupervised learning techniques.

Therefore, to apply unsupervised learning techniques, it was first necessary to extract all the data stored in the web page using the web scraping technique that allowed us to obtain all the data from the child's profile. Furthermore, since the collected data set contained inconsistencies between their records, they were preprocessed with KDD processing techniques to obtain the largest number of valid records for the study.

Finally, the data analysis was performed by varying between multiple numbers of clusters determined by the elbow method, in order to pass them to the k-means algorithm and thus determine the appropriate amount for the data set through validation metrics.

INTRODUCCIÓN

En estos años donde el uso de la tecnología es constante y tiende a un crecimiento exponencial, los datos han demostrado cumplir un rol primordial para mejorar las estrategias en cualquier rubro mediante su estudio y análisis. Se puede observar dentro de diferentes campos como el análisis de datos adecuado puede mejorar la eficiencia y eficacia de los procesos que se llevan a cabo para combatir una problemática.

Las desapariciones de menores por su parte son reguladas en el Perú bajo la Ley N.º 29685 cuyo objetivo según el diario oficial el peruano es: "Dictar medidas especiales que permitan la búsqueda, localización y protección de niños, niñas, adolescentes, personas adultas mayores y personas con discapacidad física, mental o sensorial que se encuentren desaparecidas" (pág. 442436). Además, "Se considera persona desaparecida a aquella que se encuentra ausente de su domicilio habitual, respecto del cual se desconoce su paradero" (pág. 442436).

En el año 2018 el Ministerio del Interior implemento la campaña "Te estamos buscando", cuya finalidad es distribuir notas de alertas de las personas en situación de riesgo con la finalidad de que los perfiles sean redistribuidos haciendo uso de los diferentes medios de comunicación. Esta estrategia para combatir las desapariciones de menores se basó en la Alerta Amber (America's Missing: Broadcast Emergency Response) que demostró una gran eficiencia a nivel internacional. La página web de esta campaña alberga los datos de miles de personas reportadas como desaparecidas y encontradas, dentro de sus datos se puede encontrar los atributos de edad, altura, raza, etc.

El aprendizaje automático es un campo de la Inteligencia Artificial que busca formar conocimiento a través del análisis de los datos, dentro de este campo encontramos el aprendizaje no supervisado cuyo objetivo es descubrir los patrones o estructuras de un conjunto de datos sin la necesidad de un supervisor. A su vez, el análisis de Clustering es una de las técnicas de aprendizaje no supervisado más utilizadas que busca formar clústeres de acuerdo con la similitud entre los registros de datos.

Por lo tanto, podemos utilizar las técnicas de aprendizaje no supervisado como herramienta para detectar los patrones de menores desaparecidos que existen en el conjunto de datos, tomando en cuenta los datos de los perfiles como altura, edad, genero, raza, entre otros. Se escogen estos datos porque forman son características de cada persona registrada en la página web.

ÍNDICE GENERAL

DEDICATORIA I

AGRADECIMIENTOS I

RESUMEN I

ABSTRACT III

INTRODUCCIÓN IV

ÍNDICE GENERAL V

ÍNDICE DE TABLAS VIII

ÍNDICE DE FIGURAS X

CAPÍTULO I. ASPECTOS GENERALES 1

I.1. DESCRIPCIÓN DE LA SITUACIÓN ACTUAL 1

I.2. FORMULACIÓN DEL PROBLEMA 2

I.2.1. PROBLEMA GENERAL 2

I.2.2. PROBLEMAS ESPECÍFICOS 2

I.3. OBJETIVOS 2

I.3.1. OBJETIVO GENERAL 2

I.3.2. OBJETIVOS ESPECÍFICOS 3

I.4. HIPÓTESIS 3

I.5. VARIABLES 3

I.5.1. VARIABLE DEPENDIENTE 3

I.5.2. INDICADORES DE VARIABLE DEPENDIENTE 4

I.6. JUSTIFICACIÓN 4

I.6.1. CONVENIENCIA 4

I.6.2. RELEVANCIA SOCIAL 4

I.6.3. IMPLICACIONES PRÁCTICAS 4

I.6.4. VALOR TEÓRICO 5

I.7. METODOLOGÍA 5

I.7.1. TIPO DE INVESTIGACIÓN 5

I.7.2. NIVEL DE INVESTIGACIÓN 5

I.7.3. MÉTODO DE INVESTIGACIÓN 6

I.8. MATRIZ DE CONSISTENCIA 7

CAPÍTULO II. MARCO TEÓRICO 9

II.1. ASPECTOS TEÓRICOS PERTINENTES 9

II.1.1. DATA MINING 9

II.1.2. APRENDIZAJE AUTOMÁTICO 23

II.2. ANTECEDENTES DE LA INVESTIGACIÓN 33

II.2.1. ANTECEDENTES INTERNACIONALES 33

II.2.2. ANTECEDENTES NACIONALES 37

CAPÍTULO III. METODOLOGÍA 39

III.1. TIPO DE INVESTIGACIÓN 39

III.2. DISEÑO DE LA INVESTIGACIÓN 39

III.2.1. FASE 1: RECOLECTAR DATOS 39

III.2.2. FASE 2: PRE-PROCESAMIENTO DE DATOS 40

III.2.3. FASE 3: ANÁLISIS DE CLUSTERING Y VALIDACIÓN DE RESULTADOS 41

III.2.4. FASE 4: INTERPRETACIÓN DE RESULTADOS 42

III.3. POBLACIÓN Y MUESTRA 42

III.3.1. POBLACIÓN 42

III.3.2. MUESTRA 42

III.4. INSTRUMENTOS 43

III.5. RECOLECCIÓN Y ANÁLISIS DE DATOS 43

III.5.1. TÉCNICAS DE RECOLECCIÓN DE DATOS 43

III.5.2. TÉCNICAS DE ANÁLISIS DE DATOS 49

CAPÍTULO IV. RESULTADOS 52

IV.1. ETAPA 1: RECOLECTAR DATOS 52

IV.2. ETAPA 2: PRE-PROCESAMIENTO DE DATOS 52

IV.2.1. INTEGRACIÓN DE DATOS 52

IV.2.2. LIMPIEZA DE DATOS 53

IV.2.3. TRANSFORMACIÓN DE DATOS 64

IV.2.4. REDUCCIÓN DE DATOS 64

IV.3. ETAPA 3: ANÁLISIS DE CLUSTERING Y VALIDACIÓN DE RESULTADOS 65

IV.4. FASE 4: INTERPRETACIÓN DE RESULTADOS 67

IV.4.1. DISTRIBUCIÓN DE CLÚSTERES 67

CAPÍTULO V. DISCUSIÓN 89

GLOSARIO 92

CONCLUSIONES 94

RECOMENDACIONES 95

REFERENCIAS 96

ANEXOS 98

ANEXO A: DESCRIPCIÓN DE DATOS - CLÚSTER 1 98

ANEXO B: DESCRIPCIÓN DE DATOS – CLÚSTER 2 100

ANEXO C: DESCRIPCIÓN DE DATOS – CLÚSTER 3 102

ANEXO D: DESCRIPCIÓN DE DATOS – CLÚSTER 4 104

ÍNDICE DE TABLAS

Tabla 1 *Matriz de Consistencia.* _____ *7*
Tabla 2 *Población de menores desaparecidos y encontrados.* _____ *42*
Tabla 3 *Atributos de perfil de menor desaparecido.* _____ *45*
Tabla 4 *Número de datos faltantes por atributo.* _____ *50*
Tabla 5 *Perfiles duplicados por nombre.* _____ *53*
Tabla 6 *Tipos de dato por atributo del conjunto de datos inicial.* _____ *54*
Tabla 7 *Descripción de atributos numéricos (inicial).* _____ *55*
Tabla 8 *Descripción de valores numéricos (sin valores atípicos).* _____ *56*
Tabla 9 *Valores del atributo "ojos" (inicial).* _____ *57*
Tabla 10 *Valores del atributo "ojos" (formateados).* _____ *57*
Tabla 11 *Valores del atributo "cabello" (inicial).* _____ *58*
Tabla 12 *Valores del atributo "cabello" (formateados).* _____ *58*
Tabla 13 *Valores del atributo "boca" (inicial).* _____ *59*
Tabla 14 *Valores del atributo "boca" (formateados).* _____ *59*
Tabla 15 *Valores del atributo "nariz" (inicial).* _____ *60*
Tabla 16 *Valores del atributo "nariz" (formateados).* _____ *60*
Tabla 17 *Valores del atributo "raza" (inicial).* _____ *61*
Tabla 18 *Valores del atributo "raza" (formateados).* _____ *61*
Tabla 19 *Valores del atributo "género" (formateados).* _____ *62*
Tabla 20 *Estadísticas descriptivas de atributos binarios (con valores nulos).* _____ *62*
Tabla 21 *Estadísticas descriptivas de atributos nominales (con valores nulos).* _____ *62*
Tabla 22 *Estadísticas descriptivas de atributos numéricos (sin valores nulos).* _____ *63*
Tabla 23 *Estadísticas descriptivas de atributos binarios (sin valores nulos).* _____ *63*
Tabla 24 *Estadísticas descriptivas de atributos nominales (sin valores nulos).* _____ *64*
Tabla 25 *Estadísticas descriptivas del conjunto de datos (después del preprocesamiento).* _____ *64*
Tabla 26 *Estadísticas descriptivas del conjunto de datos (redimensionado).* _____ *65*
Tabla 27 *Resultados de índices de validación.* _____ *66*
Tabla 28 *Resumen de distribución (Color de ojos x Edad).* _____ *71*
Tabla 29 *Resumen de distribución (Color de cabello x Edad).* _____ *74*
Tabla 30 *Resumen de distribución (Boca x Edad).* _____ *77*
Tabla 31 *Resumen de distribución (Nariz x Edad).* _____ *80*
Tabla 32 *Resumen de distribución (Raza x Edad).* _____ *83*
Tabla 33 *Resumen de distribución (Género x Edad).* _____ *86*
Tabla 34 *Descripción de atributos numéricos - Clúster 1.* _____ *98*
Tabla 35 *Descripción de atributos nominales - Clúster 1.* _____ *98*
Tabla 36 *Distribución de valores de género - Clúster 1.* _____ *98*

Tabla 37 *Distribución de valores de color de ojos - Clúster 1.* ____ *99*
Tabla 38 *Distribución de valores de color de cabello - Clúster 1.* ____ *99*
Tabla 39 *Distribución de valores de boca - Clúster 1.* ____ *99*
Tabla 40 *Distribución de valores de nariz - Clúster 1.* ____ *99*
Tabla 41 *Distribución de valores de raza - Clúster 1.* ____ *100*
Tabla 42 *Descripción de atributos numéricos - Clúster 2.* ____ *100*
Tabla 43 *Descripción de atributos nominales - Clúster 2.* ____ *100*
Tabla 44 *Distribución de valores de género - Clúster 2.* ____ *101*
Tabla 45 *Distribución de valores de color de ojos - Clúster 2.* ____ *101*
Tabla 46 *Distribución de valores de color de cabello - Clúster 2.* ____ *101*
Tabla 47 *Distribución de valores de boca - Clúster 2.* ____ *101*
Tabla 48 *Distribución de valores de nariz - Clúster 2.* ____ *102*
Tabla 49 *Distribución de valores de raza - Clúster 2.* ____ *102*
Tabla 50 *Descripción de atributos numéricos - Clúster 3.* ____ *102*
Tabla 51 *Descripción de atributos nominales - Clúster 3.* ____ *102*
Tabla 52 *Distribución de valores de género - Clúster 3.* ____ *103*
Tabla 53 *Distribución de valores de color de ojos - Clúster 3.* ____ *103*
Tabla 54 *Distribución de valores de color de cabello - Clúster 3.* ____ *103*
Tabla 55 *Distribución de valores de boca - Clúster 3.* ____ *103*
Tabla 56 *Distribución de valores de nariz - Clúster 3.* ____ *104*
Tabla 57 *Distribución de valores de raza - Clúster 3.* ____ *104*
Tabla 58 *Descripción de atributos numéricos - Clúster 4.* ____ *104*
Tabla 59 *Descripción de atributos nominales - Clúster 4.* ____ *104*
Tabla 60 *Distribución de valores de género - Clúster 4.* ____ *105*
Tabla 61 *Distribución de valores de color de ojos - Clúster 4.* ____ *105*
Tabla 62 *Distribución de valores de color de cabello - Clúster 4.* ____ *105*
Tabla 63 *Distribución de valores de boca - Clúster 4.* ____ *105*
Tabla 64 *Distribución de valores de nariz - Clúster 4.* ____ *106*
Tabla 65 *Distribución de valores de raza - Clúster 4.* ____ *106*

ÍNDICE DE FIGURAS

Figura 1 *Distribución de personas desparecidas en el año 2020.* _____ *1*
Figura 2 *La minería de datos como un paso en el proceso de descubrimiento de conocimiento.* _____ *10*
Figura 3 *Métodos de Discretización (Binning) para suavizado de datos.* _____ *13*
Figura 4 *Una gráfica de datos de clientes en 2-D con respecto a las ubicaciones de los clientes en una ciudad, que muestra tres grupos de datos.* _____ *14*
Figura 5 *El tamaño de la población y la inversión publicitaria de 100 ciudades diferentes se muestran como círculos de color purpura. La línea verde continua indica el primer componente principal y la línea discontinua azul indica el segundo componente principal.* _____ *21*
Figura 6 *Ejemplo de clustering del conjunto de datos iris sin etiquetas de clase.* _____ *28*
Figura 7 *Ejemplo de perfil de menor desaparecido.* _____ *44*
Figura 8 *Función inicial* _____ *46*
Figura 9 *Función Scraper* _____ *47*
Figura 10 *Función obtener perfil* _____ *48*
Figura 11 *Función guardar perfil* _____ *49*
Figura 12 *Diagrama de dispersión (edad x altura) (inicial).* _____ *55*
Figura 13 *Diagrama de dispersión (edad x altura) (sin valores atípicos).* _____ *56*
Figura 14 *Método de codo aplicado al conjunto de datos de menores desaparecidos.* _____ *66*
Figura 15 *K-means aplicado a los datos de menores desaparecidos (Datos reducidos con PCA).* _____ *67*
Figura 16 *Diagrama de dispersión (Componente principal 1 x Componente principal 2).* _____ *68*
Figura 17 *Diagrama de cajas (altura x clúster).* _____ *69*
Figura 18 *Diagrama de cajas (edad x clúster).* _____ *70*
Figura 19 *Diagrama de cajas (Color de ojos x Edad).* _____ *72*
Figura 20 *Diagrama de barras (Color de ojos x Edad).* _____ *73*
Figura 21 *Diagrama de cajas (Color de cabello x Edad).* _____ *75*
Figura 22 *Diagrama de barras (Color de cabello x Edad).* _____ *76*
Figura 23 *Diagrama de cajas (Boca x Edad).* _____ *78*
Figura 24 *Diagrama de barras (Boca x Edad).* _____ *79*
Figura 25 *Diagrama de cajas (Nariz x Edad).* _____ *81*
Figura 26 *Diagrama de barras (Nariz x Edad).* _____ *82*
Figura 27 *Diagrama de cajas (Raza x Edad).* _____ *84*
Figura 28 *Diagrama de barras (Raza x Edad).* _____ *85*
Figura 29 *Diagrama de cajas (Género x Edad).* _____ *87*
Figura 30 *Diagrama de barras (Género x Edad).* _____ *88*

CAPÍTULO I. ASPECTOS GENERALES

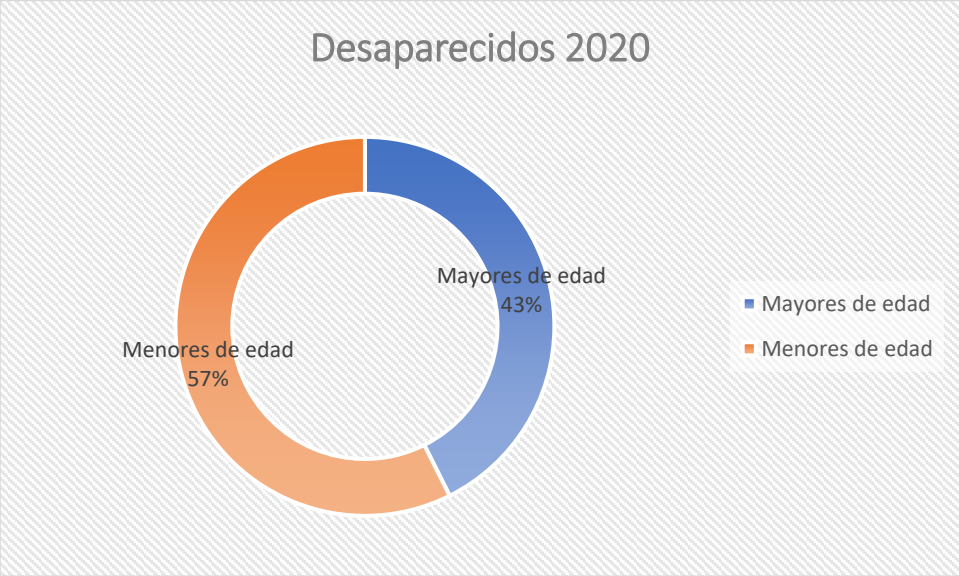
I.1. DESCRIPCIÓN DE LA SITUACIÓN ACTUAL

En el Perú en el año 2018 el Ministerio del Interior implementó la campaña "Te estamos buscando" con la finalidad de distribuir notas de alertas de personas desaparecidas que estén en condición de riesgo, para que sus perfiles sean visualizados mediante diferentes medios de comunicación.

Según el Registro Nacional de Información de Personas Desparecidas (RENIPED), el 2020 se registraron un total 18481 de personas desaparecidas, de los cuales 10605 (57.38%) eran menores de edad como se muestra en la **Figura 1**.

Figura 1

Distribución de personas desparecidas en el año 2020.



Nota. Proporción de personas desaparecidas en el año 2020 según su edad.

> ¨Si bien en octubre de 2020 se presentó el nuevo Sistema de Búsqueda de Personas Desaparecidas —que incluye el **Registro Nacional de Personas Desaparecidas (RENIPED)**, el **Portal de Personas Desaparecidas**, la **Línea 114** y el **Sistema de Mensajería de Alerta Temprana de Emergencia (SISMATE)**...Sumado a ello, dista todavía de ser un sistema interinstitucional, pues la búsqueda recae principalmente en la Policía Nacional del Perú y en el Ministerio del Interior (MININTER).¨ (Amnistía Internacional , 2021, pág. 4)

La iniciativa de la campaña mejora la distribución de información sobre los menores desaparecidos, pero no hace uso de los datos para mejorar las estrategias de búsqueda si no que plantea la acción rápida entre instituciones que como menciona Amnistía Internacional no cumple con sus fines por el momento.

Ante el problema de falta de uso de los datos de perfiles de menores desaparecidos, se plantea la construcción de un modelo para descubrir los patrones subyacentes que podrían mostrar conocimiento sobre los casos de desaparición utilizando técnicas de Minería de datos y Análisis de Clustering.

I.2. FORMULACIÓN DEL PROBLEMA

I.2.1. PROBLEMA GENERAL

¿Qué patrones se identificarán dentro del conjunto de datos de perfiles de menores desaparecidos mediante técnicas de minería de datos y análisis de clustering?

I.2.2. PROBLEMAS ESPECÍFICOS

- ¿De qué manera se puede recolectar la información de los perfiles de menores desaparecidos de la página web?
- ¿Qué atributos de los perfiles se utilizarán para aplicar las técnicas de minería de datos y clustering?
- ¿Qué estrategias de preprocesamiento se deben aplicar al conjunto de datos para realizar el análisis de clustering?
- ¿Cuál es el número de clústeres adecuado para el conjunto de datos?
- ¿Cuáles son las características de los patrones encontrados en el conjunto de datos?

I.3. OBJETIVOS

I.3.1. OBJETIVO GENERAL

Determinar los patrones existentes dentro del conjunto de datos de los perfiles de menores desaparecidos mediante técnicas de minería de datos y análisis de clustering.

I.3.2. OBJETIVOS ESPECÍFICOS

- Implementar una herramienta para recolectar el conjunto de datos de la página web de Te Estamos Buscando.
- Seleccionar los atributos del conjunto de datos que sean más relevantes para aplicar las técnicas de minería de datos y clustering.
- Seleccionar técnicas de preprocesamiento de datos previo al análisis de clustering.
- Determinar el número de clústeres adecuado para el conjunto de datos.
- Describir los clústeres determinados como óptimos para el conjunto de datos.

I.4. HIPÓTESIS

El estudio permite encontrar patrones (grupos) que comparten características dentro del conjunto de datos mediante la aplicación de técnicas de aprendizaje no supervisado con un nivel óptimo de validación.

I.5. VARIABLES

I.5.1. VARIABLE DEPENDIENTE

- Nivel de eficiencia en segmentación de patrones del conjunto de datos

Este nivel de eficiencia se refiere al resultado alcanzado por la estructura de clustering con un determinado número de clústeres seleccionado, cuyo criterio será el índice de Caliński y Harabasz y el índice de Davies-Bouldin.

Nivel de eficiencia en segmentación de patrones:

- El número de clústeres que maximice el índice de Caliński y Harabasz - $(CH(k))$ y minimice el índice de Davies-Bouldin - $DB(k)$.

Limitantes:

- El Índice de Caliński y Harabasz debe ser un valor entero positivo mayor que 0.

$$CH(k) > 0$$

- El Índice de Davies-Bouldin debe estar en el rango entre -1 y 1.

$$-1 < DB(k) > 1$$

I.5.2. INDICADORES DE VARIABLE DEPENDIENTE

- Número de clústeres.
- Índice de Caliński y Harabasz.
- Índice de Davies-Bouldin.

I.6. JUSTIFICACIÓN

I.6.1. CONVENIENCIA

Las medidas implementadas por el Ministerio de Interior ayudan a la difusión de la información primordial para ubicar a las personas desaparecidas y al fácil acceso a los perfiles de dichas personas, pero este conjunto de datos no se reutiliza con otro fin dejando un vacío de conocimiento sobre la información recolectada.

Por lo tanto, es necesario utilizar estos datos para obtener una mejor percepción sobre los casos de menores desaparecidos en el Perú. En este sentido, aplicar técnicas de aprendizaje no supervisado al conjunto de datos de los perfiles de personas desaparecidas permitirá visualizar los patrones (grupos subyacentes) que se encuentran implícitos en dicho conjunto de datos.

I.6.2. RELEVANCIA SOCIAL

Con la elaboración de esta investigación se pretende analizar los datos recolectados de los menores desaparecidos en el Perú, para así determinar si los patrones encontrados puedan ser útiles no solo para disminuir la cantidad de casos, sino también para mejorar las estrategias de búsqueda de los desaparecidos.

I.6.3. IMPLICACIONES PRÁCTICAS

Esta investigación nos ayudara a ver la importancia de los datos y su análisis. Además, nos permitirá plantear el uso de herramientas computacionales, capaces de reconocer patrones en los datos, a las entidades encargadas de combatir las desapariciones forzadas en el Perú con el fin de aplicar estrategias especializadas frente a los diferentes grupos encontrados.

I.6.4. VALOR TEÓRICO

El agrupamiento (clustering) como estrategia de aprendizaje automático no supervisado se enfoca en reconocer grupos dentro de un conjunto de datos con registros no etiquetados, es decir, que ayuda a determinar que subconjuntos dentro de un conjunto de datos están más relacionados internamente (entre sus elementos) y menos relacionados externamente (entre subconjuntos), haciendo uso de una distancia calculada entre objetos para determinar la separación en un espacio multidimensional.

Esta estrategia se aplicará al conjunto de datos de personas desaparecidas para extraer patrones, cuya información podrá ser utilizada para aplicar estrategias de respuesta específica para cada uno. Así mismo se busca nutrir las bases teóricas sobre el aprendizaje automático no supervisado y su aplicación frente a conjuntos de datos de diferente índole.

I.7. METODOLOGÍA

I.7.1. TIPO DE INVESTIGACIÓN

El tipo de investigación a realizar en este proyecto será de tipo básica. Una investigación básica se define como aquella que, "… está dirigida a un conocimiento más completo a través de la comprensión de los aspectos fundamentales de los fenómenos, de los hechos observables o de las relaciones que establecen los entes" (CIENCIACTIVA, 2016, pág. 7).

I.7.2. NIVEL DE INVESTIGACIÓN

De acuerdo con la naturaleza del proyecto de investigación está alineada con el nivel descriptivo que, "… únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, esto es, su objetivo no es indicar cómo se relacionan éstas" (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014, pág. 92).

En esta investigación se busca recolectar información de los perfiles de personas desaparecidas y a través de la aplicación de técnicas de aprendizaje no supervisado determinar la idoneidad y/o existencia de clústeres o patrones dentro del conjunto de datos para finalmente describirlos.

I.7.3. MÉTODO DE INVESTIGACIÓN

La estrategia de investigación a utilizar será la de experimento, la cual "se enfoca en investigar algunas variables y las formas en que éstas se ven afectadas por las condiciones experimentales. Típicamente, los experimentos se utilizan para verificar o falsificar una hipótesis formulada previamente" (Berndtsson, Hansson, Olsson, & Lundell, 2007, pág. 65).

El método por utilizar en este proyecto de investigación está alineado con el diseño experimental, que, según Hernández Sampieri y otros: "se utilizan cuando el investigador pretende establecer el posible efecto de una causa que se manipula" (pág. 130).

En esta investigación se pretende experimentar con los datos recolectados de la página "Te Estamos Buscando" variando el número de clústeres que se utiliza como parámetro de entrada en las técnicas de clustering para así comparar los resultados de los índices de validación, dicho enfoque se inspira en las técnicas de validación relativa de clustering. Esto permitirá determinar el número de patrones que existen de forma natural dentro del conjunto de datos.

I.8. MATRIZ DE CONSISTENCIA

Tabla 1

Matriz de Consistencia.

Problema	Objetivo	Hipótesis	Metodología
General ¿Qué patrones se identificarán dentro del conjunto de datos de perfiles de menores desaparecidos mediante técnicas de minería de datos y análisis de clustering? Específicos ¿De qué manera se puede recolectar la información de los perfiles de menores desaparecidos de la página web? ¿Qué atributos de los perfiles se utilizarán para aplicar las técnicas de minería de datos y clustering? ¿Qué estrategias de preprocesamiento de deben aplicar al conjunto de datos para realizar el análisis de clustering? ¿Qué estrategias de preprocesamiento	General Determinar los patrones existentes dentro del conjunto de datos de los perfiles de menores desaparecidos mediante técnicas de minería de datos y análisis de clustering. Específicos Implementar una herramienta para recolectar el conjunto de datos de la página web de “Te Estamos Buscando”. Seleccionar los atributos del conjunto de datos que sean más relevantes para aplicar las técnicas de minería de datos y clustering. Seleccionar técnicas de preprocesamiento de datos previo al análisis de clustering.	General El estudio permite encontrar patrones que comparten características dentro del conjunto de datos mediante la aplicación de técnicas de aprendizaje no supervisado con un nivel óptimo de validación. Nula El estudio no permite encontrar patrones que comparten características dentro del conjunto de datos mediante la aplicación de técnicas de aprendizaje no supervisado con un nivel óptimo de validación.	Tipo Investigación de tipo básica Nivel Descriptivo Método Experimental

de deben aplicar al conjunto de datos para realizar el análisis de clustering? ¿Cuál es el número de clústeres adecuado para el conjunto de datos? ¿Cuáles son las características de los patrones encontrados en el conjunto de datos?	Determinar el número de clústeres adecuado para el conjunto de datos. Describir los clústeres determinados como óptimos para el conjunto de datos.

CAPÍTULO II. MARCO TEÓRICO

II.1. ASPECTOS TEÓRICOS PERTINENTES

II.1.1. DATA MINING

Es un campo derivado de la estadística y las ciencias de la computación que busca generar información primordial que se convierte en conocimiento esencial a partir de una base de datos.

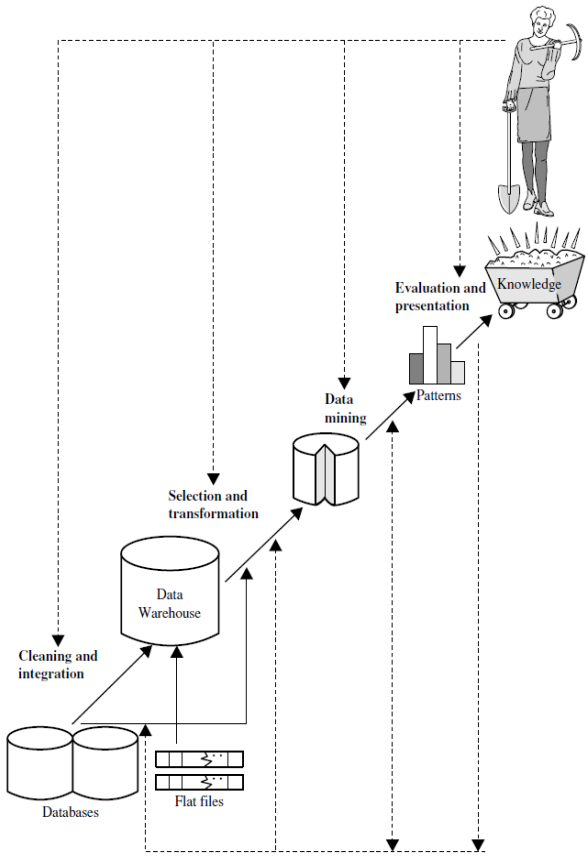
> Muchas personas tratan la minería de datos como sinónimo de otro término utilizado popularmente, descubrimiento de conocimiento a partir de datos o KDD (Knowledge discovery from data), mientras que otros ven la minería de datos como un simple paso esencial en el proceso de descubrimiento de conocimiento. (Han, Kamber, & Pei, 2011, pág. 6)

El descubrimiento de conocimiento se da como una secuencia iterativa de los siguientes pasos:

1. **Limpieza de datos:** remover ruido y datos inconsistentes.
2. **Integración de datos:** combinar múltiples fuentes de datos.
3. **Selección de datos:** recuperar datos relevantes para el análisis.
4. **Transformación de datos:** transformar datos en formas adecuadas para su procesamiento.
5. **Procesamiento de datos (Data mining):** aplicar métodos inteligentes para extraer patrones de datos.
6. **Evaluación de patrones:** Identificar los patrones interesantes.
7. **Presentación de conocimientos:** Presentar el conocimiento minado.

Figura 2

La minería de datos como un paso en el proceso de descubrimiento de conocimiento.



Nota. Pasos para obtener conocimiento de los datos, del paso 1 al 4 son diferentes formas de preprocesamiento de datos. Fuente: (Han, Kamber, & Pei, 2011, pág. 7).

II.1.1.1. PREPROCESAMIENTO DE DATOS

En la sección 2.1.1 hablamos de la secuencia de pasos que se deben de realizar para minar conocimiento de los datos. Una parte de estos pasos pertenecen a lo que llamamos el preprocesamiento de los datos que consiste en técnicas utilizadas para mejorar la calidad de nuestro conjunto de datos debido a que, "los datos de baja calidad conducirán a resultados de minería de baja calidad" (Han, Kamber, & Pei, 2011, pág. 83).

Existen múltiples técnicas de preprocesamiento que podemos aplicar, estas son:

1. Limpieza de datos.
2. Integración de datos.
3. Reducción de datos.

4. Transformación de datos.

Estas técnicas se complementan entre sí para mejorar la calidad del conjunto de datos y su resultado puede "… mejorar la precisión y la eficiencia de los algoritmos de minería que involucran mediciones de distancia. Estas técnicas no se excluyen mutuamente; pueden trabajar juntos" (Han, Kamber, & Pei, 2011, pág. 83).

II.1.1.1.1. LIMPIEZA DE DATOS

> Los datos del mundo real tienden a ser incompletos, ruidosos e inconsistentes. Las rutinas de limpieza de datos (o limpieza de datos) intentan completar los valores faltantes, suavizar el ruido al identificar valores atípicos y corregir inconsistencias en los datos. (Han, Kamber, & Pei, 2011, pág. 88)

Para limpiar los datos se debe estudiar múltiples métodos básicos, dentro de estos encontramos métodos o técnicas para tratar los valores faltantes y el suavizado de datos.

II.1.1.1.1.1 VALORES FALTANTES

Típicamente podemos encontrar conjuntos de datos con valores faltantes o nulos, que pueden representar un mal almacenamiento o registro de los datos, o también esta situación se puede dar debido a que ciertos atributos de nuestros objetos de datos consideran valida la existencia y ausencia de los valores de algunos atributos (por ejemplo: *los datos binarios*).

Para completar valores faltantes existen diferentes métodos, los cuales son:

1. **Ignorar la tupla:** Consiste en ignorar los atributos de las tuplas con el valor faltante, esto se realiza generalmente cuando falta la etiqueta de la clase en problemas de clasificación, "Este método no es muy eficaz, a menos que la tupla contenga varios atributos con valores perdidos" (Han, Kamber, & Pei, 2011, pág. 88).

2. **Complete el valor faltante manualmente:** Este método puede no ser conveniente, ya que consume mucho tiempo.

3. **Utilice una constante global para llenar el valor faltante:** Este puede ser un método simple, pero al utilizar el mismo valor para todos los atributos puede confundir al algoritmo y ser falible.

4. **Utilice una medida de tendencia central para el atributo para completar el valor que falta:** "Para distribuciones de datos normales (simétricas), se puede usar la media, mientras que la distribución de datos asimétrica debe emplear la mediana" (Han, Kamber, & Pei, 2011, pág. 88).

5. **Utilice el atributo medio o mediana para todas las muestras que pertenezcan a la misma clase que la tupla dada:** Igualmente que el anterior método, dependiendo a la distribución de los datos podemos variar entre media y mediana.

6. **Utilice el valor más probable para completar el valor que falta:** "Esto puede determinarse con regresión, herramientas basadas en inferencia utilizando un formalismo bayesiano o inducción de árbol de decisión" (Han, Kamber, & Pei, 2011, págs. 88,89).

La mayoría de estos métodos sesgan los datos. Sin embargo, el último método mencionado es el más popular debido a que utiliza todos los datos posibles del conjunto para completar los valores faltantes.

II.1.1.1.1.2 DATOS RUIDOSOS

El ruido en los datos se define como "… un error aleatorio o una variación en una variable medida" (Han, Kamber, & Pei, 2011, pág. 89). Podemos utilizar métodos de visualización de datos para detectar valores atípicos o ruido en los atributos del conjunto de datos, la estrategia encargada de remover el ruido de los datos se denomina suavizado de datos. Dentro de esta existen técnicas como:

1. **Discretización (Binning):** Suavizan un valor de datos ordenados observando los valores que lo rodean, los valores se distribuyen en múltiples contenedores. "Dado que los métodos de agrupación consultan la vecindad de valores, realizan un suavizado local" (Han, Kamber, & Pei, 2011, pág. 89). Cada contenedor tiene la misma cantidad de valores (o frecuencia) y para suavizarlo se puede realizar el reemplazo de valores usando la media, la mediana o los límites del contenedor, este último consiste en reemplazar los valores dentro de un contenedor con el límite más cercano el valor.

Figura 3

Métodos de Discretización (Binning) para suavizado de datos.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

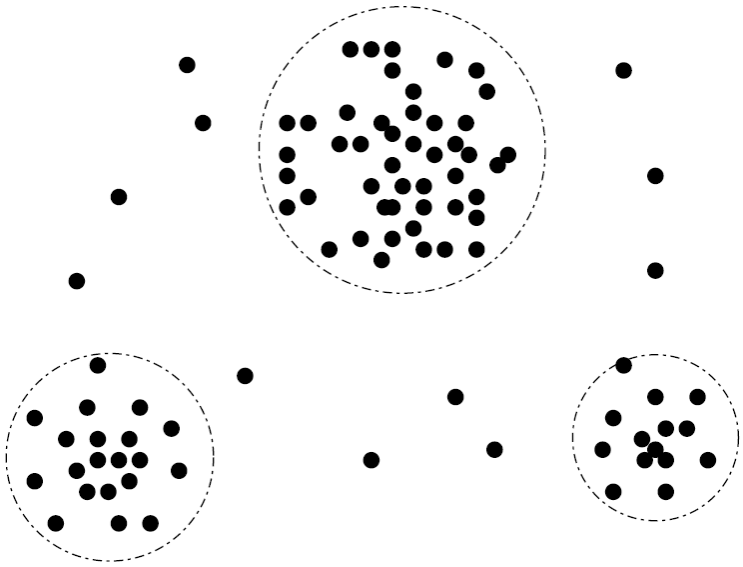
Nota. Fuente: (Han, Kamber, & Pei, 2011, pág. 90)

2. **Regresión:** "El suavizado de datos también se puede realizar mediante regresión, una técnica que adapta los valores de los datos a una función" (Han, Kamber, & Pei, 2011, pág. 90). Existen regresión lineal (2 dimensiones) y regresión lineal múltiple (3 o más dimensiones), esta técnica busca encontrar la mejor línea o figura (dependiendo de las dimensiones) a la que los valores se ajusten para así poder predecir el valor de una dimensión usando otro atributo.

3. **Análisis de valores atípicos:** los valores atípicos pueden ser detectados por métodos visuales o por medio de clustering, así cualquier valor que este fuera de los clústeres se consideran atípicos.

Figura 4

Una gráfica de datos de clientes en 2-D con respecto a las ubicaciones de los clientes en una ciudad, que muestra tres grupos de datos.



Los valores atípicos pueden detectarse como valores que quedan fuera de los conjuntos de grupos. Fuente: (Han, Kamber, & Pei, 2011, pág. 91).

Varios métodos de suavizado de datos pueden ser utilizados para transformación o reducción de datos, como es el caso de la discretización (Binning) que reduce la cantidad de valores distintos dentro de un conjunto de datos.

II.1.1.1.2. INTEGRACIÓN DE DATOS

La integración de datos consiste en "la fusión de datos de varios almacenes de datos" (Han, Kamber, & Pei, 2011, pág. 93). Este proceso puede convertirse en uno muy complejo dependiendo a las características de los datos y de las fuentes de donde se extraen, por esta razón se presentan diferentes problemas en la integración de datos, entre estos están:

II.1.1.1.2.1 PROBLEMA DE IDENTIFICACIÓN DE ENTIDADES.

Este problema se centra en identificar que atributos que corresponden o están relacionados entre diferentes conjuntos de datos; esto se debe realizar previo a la integración de datos, debido a que pueden existir incongruencias en las características de los atributos causando así una mala integración de datos.

Para solucionar este problema se debe realizar una comparación de las entidades albergadas en las diferentes fuentes de datos que se utilizaran, es necesario analizar los conocimientos previos que tenemos de cada atributo, conocidos como "metadatos" dentro de los cuales están las características de los atributos como el nombre, el significado, el tipo de dato, el rango de valores permitidos y las reglas que se aplican a cada uno.

II.1.1.1.2.2 ANALISÍS DE REDUNDANCIA Y CORRELACIÓN.

Después de realizar la integración de datos podemos encontrarnos con problemas de redundancia, debido a que se integran múltiples conjuntos de datos. "Algunas redundancias pueden detectarse mediante análisis de correlación" (Han, Kamber, & Pei, 2011, pág. 94).

El análisis de correlación se realiza tomando dos atributos y se determina la técnica a utilizar según los tipos de datos, entre estas técnicas están la de chi-cuadrado para datos nominales y coeficiente de correlación y la covarianza para datos numéricos; con estas técnicas podemos medir que tanto influye una variable en la otra basándonos en los registros actuales.

A. Prueba de correlación para datos nominales (chi-cuadrado).

Dados dos atributos A y B, la prueba de chi-cuadrado (X^2) se calcula como:

$$X^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Donde:

- c: número de columnas.
- r: número de filas.
- o_{ij}: frecuencia observada.
- e_{ij}: frecuencia esperada.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Donde:

- n: número de tuplas de datos.
- $count(A = a_i)$: número de tuplas que tienen el valor a_i para A.
- $count(B = b_j)$: número de tuplas que tienen el valor de b_j para B.

El estadístico de chi-cuadrado "prueba la hipótesis de que A y B son independientes, es decir, no existe correlación entre ellos. La prueba se basa en un nivel de significancia, con $(r - 1) \times (c - 1)$ grados de libertad" (Han, Kamber, & Pei, 2011, pág. 95).

B. Coeficiente de correlación para datos numéricos.

Dados dos atributos A y B, el coeficiente de correlación se calcula como:

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

Donde:

- n: número de tuplas.
- a_i, b_i: valor de A, B en la tupla i.
- $\bar{A}, \bar{B}$: medias de A, B.
- σ_A, σ_B: las desviaciones estándar de A, B.
- $\sum_{i=1}^{n}(a_i b_i)$: la suma del producto cruz de AB.

Se debe tener en cuenta que: $-1 \leq r_{A,B} \leq +1$. Si el valor resultante es mayor que 0 indica una correlación positiva, si es menor que 0 indica una correlación negativa y si el valor es igual a 0 indica que los atributos son independientes.

C. Covarianza de datos numéricos.

Dados dos atributos A y B, los valores esperados (media) se calculan como:

$$E(A) = \bar{A} = \sum_{i=1}^{n} a_i \wedge E(B) = \bar{B} = \sum_{i=1}^{n} b_i$$

Donde: n – número de tuplas.

Y la covarianza entre A y B se define como:

$$Cov(A,B) = E\big((A - \bar{A})(B - \bar{B})\big) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

Si comparamos las dos últimas ecuaciones (covarianza y coeficiente de correlación), vemos que:

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

II.1.1.1.2.3 DUPLICACIÓN DE TUPLAS.

Discrepancias surgen debido a la duplicidad de tuplas que se puede dar por una entrada de datos inexacta o por la falta de actualización de algunos registros. Según Han, Kamber y Pei: "Además de detectar redundancias entre atributos, la duplicación también debe detectarse a nivel de tupla" (pág. 98).

II.1.1.1.2.4 DETECCIÓN Y RESOLUCIÓN DE CONFLICTOS DE VALOR DE DATOS

Cuando se integran múltiples conjuntos de datos surgen conflictos, como los mencionados en los puntos anteriores por esta razón el proceso de integración de datos "… también involucra la detección y resolución de conflictos de valor de datos. … Esto puede deberse a diferencias en la representación, escala o codificación" (Han, Kamber, & Pei, 2011, pág. 99).

II.1.1.1.3. REDUCCIÓN DE DATOS

Debido a las características de los conjuntos de datos que se manejan en la actualidad es común encontrarnos con múltiples inconsistencias; debido a esto, además de realizar un análisis visual de los registros es necesario aplicar técnicas que ayuden a mejorar la relevancia de los datos y reducir su complejidad de procesamiento.

> Se pueden aplicar técnicas de reducción de datos para obtener una representación reducida del conjunto de datos que es mucho más pequeño en volumen, pero que mantiene fielmente la integridad de los datos originales. Es decir, la minería en el conjunto de datos reducido debería ser más eficiente, pero producir los mismos (o casi los mismos) resultados analíticos. (Han, Kamber, & Pei, 2011, pág. 99)

Para realizar este proceso podemos aplicar diferentes estrategias, entre las cuales están:

- **Reducción de dimensionalidad:** Según Han, Kamber y Pei (2011): "… es el proceso de reducir la cantidad de variables aleatorias o atributos bajo consideración" (pág. 99). Entre los métodos de esta estrategia podemos encontrar: transformaciones de ondículas (Wavelet Transforms), análisis de componentes principales (Principal Components Analysis) y selección de subconjuntos de atributos. Además, "… es una técnica popular para eliminar atributos ruidosos (es decir, irrelevantes) y redundantes (también conocidos como características)" (Aggarwal & Reddy, 2014, pág. 30).

- **Reducción de numerosidad:** Es el proceso en el cual se utilizan técnicas que "… reemplazan el volumen de datos original por formas alternativas más pequeñas de representación de datos" (Han, Kamber, & Pei, 2011, pág. 100). Las técnicas de esta estrategia pueden ser paramétricas o no paramétricas, las paramétricas solo necesitan guardar los parámetros utilizados para reproducir los datos en lugar del conjunto de datos y, por otro lado, entre las técnicas no paramétricas podemos encontrar histogramas, agrupamiento, muestreo y agregación de cubos de datos.

- **Compresión de datos:** En esta estrategia "… se aplican transformaciones para obtener una representación reducida o *comprimida* de los datos originales" (Han, Kamber, & Pei, 2011, pág. 100). Si los datos originales pueden ser reconstruidos en su totalidad sin pérdida de información a partir de los datos comprimidos, se le denomina *sin perdida*. Por otro lado, si solo podemos reconstruir una aproximación de los datos originales se denomina *con perdida*.

II.1.1.1.3.1 ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

Cuando se manejan conjuntos de datos de alta dimensionalidad con atributos correlacionados, "los componentes principales nos permiten resumir este conjunto con un número menor de variables representativas que explican colectivamente la mayor parte de la variabilidad en el conjunto original" (James, Witten, Hastie, & Tibshirani, 2013, pág. 374). Este proceso de análisis comprende el cálculo de los componentes principales y su uso para comprender los datos.

Además, nos sirve como herramienta para la visualización de datos, debido a que los conjuntos de alta dimensionalidad no pueden ser representados adecuadamente en todas sus dimensiones o, por otro lado, tendríamos que realizar $p(p-1)$ diagramas de dispersión en dos dimensiones, donde: d es el número de características o dimensiones del conjunto de datos, cuya dispersión de dimensiones no permitiría el análisis visual adecuado de los registros. Por lo tanto, un conjunto considerando solo los componentes principales, que no son más que combinaciones lineales de las dimensiones iniciales, nos permitiría representar el mayor porcentaje de características de los datos en un número de dimensiones reducido.

Los componentes principales o las nuevas dimensiones reducidas se calculan de la siguiente manera: El primer componente principal de un conjunto de características $X_1, X_2, \dots, X_p$ es la combinación lineal normalizada de estas:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{d1}X_p$$

En este componente se contiene la mayor cantidad de varianza del conjunto de datos en todas sus dimensiones y es normalizado debido a que cada dimensión tiene una carga tal que $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Juntas las cargas conforman el vector de carga del componente principal, $\phi_1 = \left(\phi_{11}\ \phi_{21}\ \dots\ \phi_{p1}\right)^T$.

Dado un conjunto de datos X de $n \times p$, para calcular el componente principal primero cada registro del conjunto X debe tener una media de cero. Luego se calcula la combinación lineal de los valores de cada dimensión:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}$$

"En otras palabras, el primer vector de carga del componente principal resuelve el problema de optimización" (James, Witten, Hastie, & Tibshirani, 2013, p. 376).

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \; sujeto\ a \sum_{j=1}^{d} \phi_{j1}^2 = 1$$

Por lo tanto, el objetivo de la función será $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$. Dado que $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, el promedio de $z_{11}, \ldots, z_{n1}$ también será cero.

> Hay una buena interpretación geométrica para el primer componente principal. El vector de carga ϕ_1 con elementos $\phi_{11}, \phi_{21,}, \ldots, \phi_{p1}$ define una dirección en el espacio de características a lo largo de la cual los datos varían más. Si proyectamos los n puntos de datos $x_1, \ldots, x_n$ en esta dirección, los valores proyectados son las puntuaciones de los componentes principales $z_{11}, \ldots, z_{n1}$ ellos mismos. (James, Witten, Hastie, & Tibshirani, 2013, p. 376)

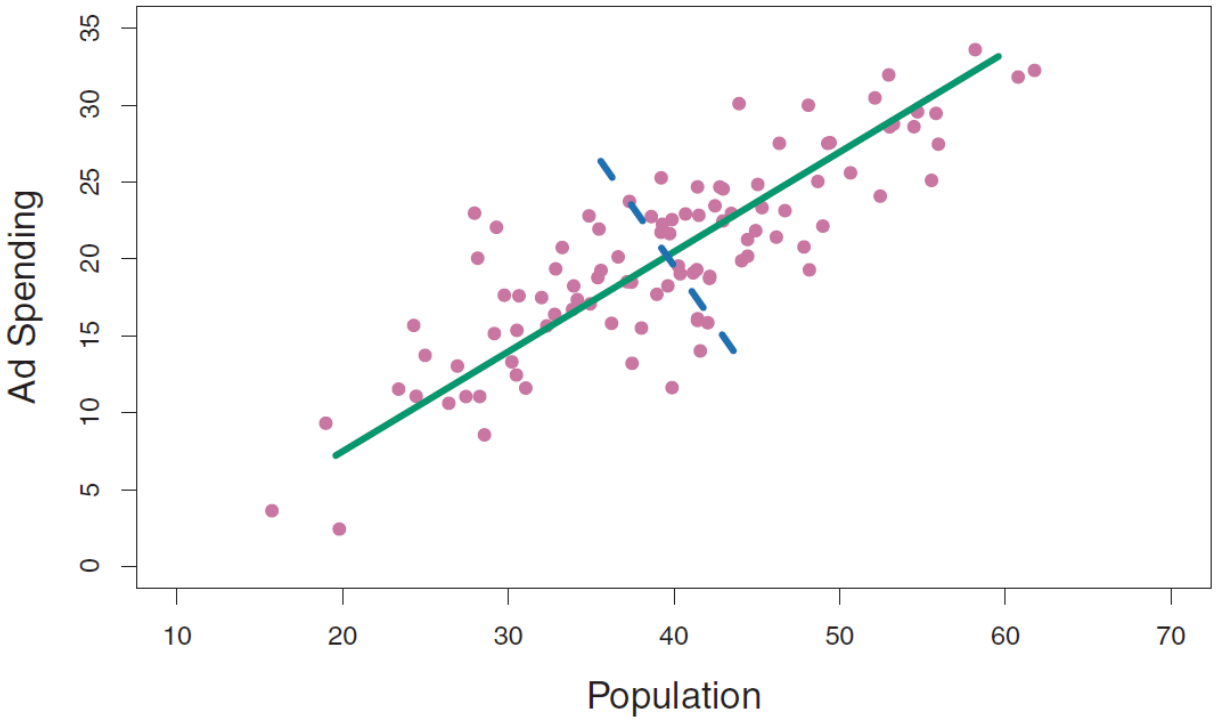
Después de calcular el primer componente principal Z_1 de las características; procedemos con el cálculo del segundo componente principal Z_2, este es la combinación lineal de $X_1, \ldots, X_p$ que tiene la máxima varianza de todas las combinaciones lineales que no están relacionadas con Z_1. Las puntuaciones del segundo componente principal $z_{12}, \ldots, z_{n2}$ toman la forma:

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip}$$

Donde ϕ_2 es el segundo vector de carga del componente principal conformado por $\phi_{11}, \phi_{21,}, \ldots, \phi_{p1}$. Para evitar que Z_1 y Z_2 estén relacionados la dirección de ϕ_2 tiene que ser ortonormal (perpendicular) a la dirección de ϕ_1.

Figura 5

El tamaño de la población y la inversión publicitaria de 100 ciudades diferentes se muestran como círculos de color purpura. La línea verde continua indica el primer componente principal y la línea discontinua azul indica el segundo componente principal.



Nota. Fuente: (James, Witten, Hastie, & Tibshirani, 2013, pág. 230)

II.1.1.1.4. TRANSFORMACIÓN DE DATOS

"En este paso de preprocesamiento, los datos se transforman o consolidan para que el proceso de minería resultante sea más eficiente y los patrones encontrados sean más fáciles de entender" (Han, Kamber, & Pei, 2011, pág. 111).

Dentro de los métodos de transformación de datos encontramos:

1. **Suavizado:** Elimina el ruido de los datos.
2. **Construcción de atributos:** Crea nuevos atributos o características para facilitar el proceso de minería.
3. **Agregación:** Aplica operaciones a los datos para acumularlos o extraer un resumen según parámetros especificados.

4. **Normalización:** Escala los datos para ponerlos dentro de un rango de valores más pequeño.

5. **Discretización:** Cambia los valores numéricos por etiquetas de conceptos o rangos, posteriormente se puede organizar recursivamente dichas etiquetas para crear una jerarquización de conceptos de nivel superior para el valor numérico.

6. **Generación de jerarquías para conceptos de datos nominales:** Cambia los valores de atributos nominales para generalizarlos en conceptos de nivel superior.

II.1.1.1.4.1 TRANSFORMACIÓN DE ATRIBUTOS DISCRETOS A NUMÉRICOS

La mayoría de los algoritmos de aprendizaje automático están enfocados en manejar datos numéricos, como es el caso de k-means enfocado a utilizar las medias del conjunto de datos para determinar un centro de clúster y también realiza el cálculo de la distancia tomando en cuenta estrategias (como la distancia Euclidiana) que sean eficientes.

Pero al tratarse dentro de un contexto distinto como son los datos con atributos mixtos (numéricos y nominales), se busca estrategias de conversión que ajusten los datos nominales a numéricos.

Una técnica sugiere dar un valor numérico que represente las posibles variantes del atributo nominal, por ejemplo: *Rojo=0, Azul=1 y Verde=2* pueden ser considerados como valores nominales y su asignación numérica respectivamente, pero no representaría una codificación adecuada, debido a que la distancia calculada es diferente entre cada uno de los valores. Esta técnica podría utilizarse con los datos de tipo nominal ordinal.

Por otro lado, Witten, Frank y Hall recomiendan que se “reemplace un atributo nominal con valor k por k atributos binarios sintéticos, uno por cada valor que indica si el atributo tiene ese valor o no” (pág. 322), esto mejora la precisión de la distancia o similitud calculada entre los registros.

II.1.2. APRENDIZAJE AUTOMÁTICO

El aprendizaje automático, aprendizaje de maquina o "Machine Learning", también denominado aprendizaje estadístico, forma parte del campo de la Inteligencia Artificial, se enfoca en obtener conocimiento a partir del análisis de los datos y transferirlo a la máquina. Además, Alpaydin menciona que "El aprendizaje automático consiste en programar computadoras para optimizar un criterio de rendimiento utilizando datos de ejemplo o experiencias pasadas" (pág. 3).

Por lo tanto, el termino de "aprender" o "aprendizaje" en este contexto no tiene una definición concreta, según Witten, Ian H. y otros "Las cosas aprenden cuando cambian su comportamiento de una manera que los hace desempeñarse mejor en el futuro" (pág. 7). También se define que "El aprendizaje implica pensamiento y propósito. Algo que aprende tiene que hacerlo de forma intencionada. … Aprender sin propósito es simplemente entrenamiento" (Witten, Frank, & Hall, 2005). Entonces, el aprendizaje que es parte de esta teoría se define mejor como entrenamiento vinculándolo así con el desempeño en lugar del conocimiento.

Dentro del campo del aprendizaje automático, podemos disgregar o clasificar tres tipos: aprendizaje supervisado, aprendizaje no supervisado y el semi-supervisado.

II.1.2.1. APRENDIZAJE SUPERVISADO

También conocido como aprendizaje de clasificación, obtiene conocimiento de datos etiquetados para poder formular una función que pueda predecir los futuros casos de datos sin etiquetar. Es decir, este aprendizaje tiene un esquema que "… se presenta con un conjunto de ejemplos clasificados de los que se espera aprender una forma de clasificar ejemplos no vistos" (Witten, Frank, & Hall, 2005, pág. 40).

Este aprendizaje puede subdividirse a la vez en regresión o clasificación de acuerdo con la etiqueta de respuesta. "Si las etiquetas son discretas, el problema de aprendizaje se llama problema de clasificación, porque los patrones se asignan a las clases... Si las etiquetas son continuas, la tarea es un problema de regresión" (Kramer, 2013, pág. 3).

En los problemas de clasificación "… el objetivo es predecir una etiqueta de clase, que es una elección de una lista predefinida de posibilidades" (Müller & Guido, 2016, pág. 25). Por otro lado, en regresión "… el objetivo es predecir un número continuo o un número de punto flotante en términos de programación (o un número real en términos matemáticos)" (Müller & Guido, 2016, pág. 26).

II.1.2.2. APRENDIZAJE SEMI-SUPERVISADO

Denominamos aprendizaje semi-supervisado a aquel se ubica en un punto medio entre el aprendizaje supervisado y no supervisado, el cual "... se refiere al caso en el que se aprende una función de predicción en ejemplos de entrenamiento etiquetados y no etiquetados" (Amini & Usunier, 2015, pág. 33). Además, esta estrategia de aprendizaje busca usar métodos "que amplifican pequeñas cantidades de datos de entrenamiento etiquetados en más" (Skiena, 2017, pág. 374).

Debido al costo de la elaboración de un conjunto de datos etiquetados, el aprendizaje semi-supervisado es una opción viable cuyas características han demostrado una brecha corta con el aprendizaje supervisado en términos de desempeño. Según Theodoridis & Koutroumbas: "el aprendizaje semi-supervisado está ganando importancia en los últimos años y actualmente se encuentra entre las áreas de investigación más candentes" (pág. 568).

II.1.2.3. APRENDIZAJE NO SUPERVISADO

> En el aprendizaje supervisado, el objetivo es aprender una asignación de la entrada a una salida cuyos valores correctos son proporcionados por un supervisor. En el aprendizaje no supervisado, no existe tal supervisor y solo tenemos datos de entrada. El objetivo es encontrar las regularidades en la entrada. El espacio de entrada tiene una estructura tal que ciertos patrones ocurren con más frecuencia que otros, y queremos ver qué sucede generalmente y qué no. (Alpaydin, 2009, pág. 11)

> El aprendizaje no supervisado encuentra estructuras en los datos. Las etiquetas para las instancias de datos u otras formas de orientación para la capacitación no son necesarias. Esto hace que el aprendizaje no supervisado sea atractivo en aplicaciones donde los datos son baratos de obtener, pero las etiquetas son caras o no están disponibles. (Wittek, 2014, pág. 57)

Un algoritmo enfocado a este tipo de aprendizaje automatizado debe ser capaz de identificar estructuras utilizando solo los registros de los datos. Según Alpaydin el aprendizaje no supervisado: "En estadística, es llamado estimación de densidad. … Un método para la estimación de densidad es el clustering donde el objetivo es encontrar grupos o agrupaciones de entrada" (pág. 11).

Podemos definir entonces al aprendizaje no supervisado como una técnica que busca descubrir información oculta dentro de los datos para generar conocimiento. Además, James y otros mencionan que "El aprendizaje no supervisado a menudo se realiza como parte de un análisis de datos exploratorio" (p. 374).

II.1.2.3.1. CLUSTERING

"Es el proceso de encontrar grupos significativos en los datos. En clustering, el objetivo no es predecir una variable de clase objetivo, sino simplemente capturar las posibles agrupaciones naturales en los datos" (Kotu & Deshpande, 2018, pág. 221).

> El análisis de clustering o simplemente clustering es el proceso de particionar un conjunto de objetos de datos (u observaciones) en subconjuntos. Cada subconjunto es un clúster, de modo que los objetos en un clúster son similares entre sí, pero diferentes a los objetos en otros clústeres. El conjunto de clústeres resultantes de un análisis de clúster puede denominarse clustering. En este contexto, diferentes métodos de clustering pueden generar diferentes agrupaciones en el mismo conjunto de datos. La partición no es realizada por humanos, sino por el algoritmo de clustering. Por lo tanto, clustering es útil porque puede conducir al descubrimiento de grupos dentro de los datos. (Han, Kamber, & Pei, 2011, pág. 444)

> Como función de minería de datos, el clustering se puede utilizar como una herramienta independiente para obtener información sobre la distribución de datos, observar las características de cada clúster y enfocarse en un conjunto particular de clústeres para un análisis posterior. Alternativamente, puede servir como un paso de preprocesamiento para otros algoritmos, como caracterización, selección de subconjuntos de atributos y clasificación, que luego operarían en los grupos detectados y los atributos o características seleccionados. (Han, Kamber, & Pei, 2011, pág. 445)

II.1.2.3.1.1 CLUSTERING PARA DESCRIBIR LOS DATOS

"La aplicación más común de clustering es explorar los datos y encontrar todos los grupos significativos posibles en los datos" (Kotu & Deshpande, 2018, pág. 221).

Algunas de las aplicaciones de clustering para describir datos son:

- **Marketing:** Encontrar grupos de clientes basados en sus comportamientos previos, atributos de los clientes potenciales y patrones de compra. Esto es útil para ajustar el mensaje de marketing a los diferentes grupos de clientes.
- **Clustering de documentos:** Este provee una forma de identificar los temas clave, comprender y resumir estos grupos en vez de leer documentos completos.
- **Agrupación de sesiones:** En la analítica web el clustering es útil para detectar los patrones de comportamiento de los usuarios dentro de una página web, mediante el almacenamiento de la transmisión de sus acciones.

II.1.2.3.1.2 CLUSTERING PARA PREPROCESAMIENTO

> Dado que los procesos de clustering consideran todos los atributos del conjunto de datos y "reducen" la información a un clúster, que es realmente otro atributo, el clustering puede usarse como una técnica de compresión de datos. El resultado del clustering es el nombre del grupo para cada registro y se puede usar como una variable de entrada para otras tareas de minería de datos predictivos. Por lo tanto, el clustering puede emplearse como una técnica de preprocesamiento para otros procesos de minería de datos. (Kotu & Deshpande, 2018, pág. 222)

Este puede ser usado para dos tipos de preprocesamiento:

- **Clustering para reducir dimensionalidad:** En todo conjunto de datos con una determinada cantidad de dimensiones existe una complejidad de cálculo proporcional al número de dimensiones. Con el clustering podemos reducir la dimensionalidad de un conjunto de datos a un atributo categórico reduciendo así la complejidad, aunque también involucrara pérdida de información.

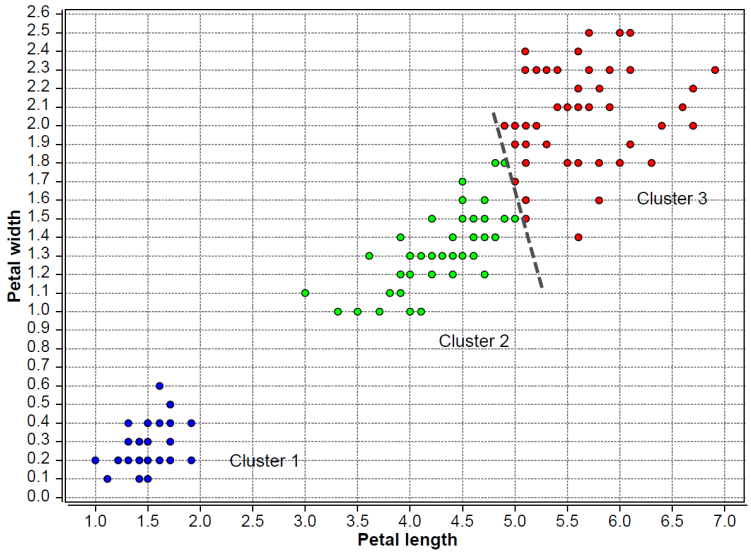
- **Clustering para reducción de objetos:** Mediante el clustering podemos formar clústeres dentro del conjunto de datos, con lo cual podemos reducir la cantidad de objetos a prototipos de los clústeres encontrados, cuyos atributos sean los más representativos de los clústeres. Finalmente podríamos utilizar estos prototipos para realizar un análisis con algoritmos de regresión o clasificación, lo que conllevaría a una reducción en el tiempo de procesamiento.

II.1.2.3.1.2.1 TIPOS DE TECNICAS DE CLUSTERING

Sin importar los tipos de técnicas de clustering, el objetivo de todos es encontrar los subgrupos del conjunto de datos, de tal forma que los objetos que se encuentren dentro de un clúster sean más similares entre sí que a los objetos de otros clústeres. “Una de las formas comunes de medir la similitud es la medición de distancia euclidiana en el espacio n-dimensional que se utiliza en muchos algoritmos de agrupamiento” (Kotu & Deshpande, 2018, pág. 223).

Figura 6

Ejemplo de clustering del conjunto de datos iris sin etiquetas de clase.



Nota. Fuente: (Kotu & Deshpande, 2018, pág. 220).

Antes de proceder con la clasificación de las técnicas de clustering debemos tener en cuenta como se distribuyen los clústeres de acuerdo con la pertenencia de sus objetos, según esto los clústeres se dividen en:

- **Clústeres de partición exclusiva o estricta:** Donde cada registro pertenece a un único clúster.
- **Clústeres superpuestos:** Los clústeres no son exclusivos y un registro puede pertenecer a más de uno.
- **Clústeres jerárquicos:** Los clústeres pueden dividirse o aglomerarse y existen clústeres padres e hijos.
- **Clústeres difusos o probabilísticos:** Donde cada registro es parte de todos los clústeres variando en un grado de pertenencia entre 0 y 1.

> Las técnicas de clustering también se pueden clasificar en función del enfoque algorítmico utilizado para encontrar clústeres en el conjunto de datos. Cada una de estas clases de algoritmos de agrupación difiere según la relación que aprovechan entre los objetos de datos. (Kotu & Deshpande, 2018, pág. 224)

- **Clustering basado en prototipos:** Donde cada clúster es representado por un objeto central, llamado prototipo, que a menudo es el centro de dicho clúster por lo cual también es denominado como clustering basado en el centro.

- **Clustering de densidad:** Donde los clústeres son denominados de acuerdo con la densidad de los objetos en el espacio y son rodeados por áreas de objetos de baja densidad.

- **Clustering jerárquico:** Es un proceso donde se crea una jerarquía de clústeres de acuerdo con la distancia entre sus puntos. El resultado de este tipo es un dendrograma, el cual es un diagrama de árbol donde se puede observar diferentes clústeres de acuerdo con un punto de precisión. Hay dos enfoques para crear una jerarquía de clústeres: aglomerativo (de abajo hace arriba) y divisivo (de arriba hace abajo).

- **Clustering basado en modelos:** Esta basado en la estadística y los modelos de distribución de probabilidad, en este los clústeres pueden ser vistos como agrupaciones que tienen los objetos pertenecientes a una misma distribución de probabilidad.

II.1.2.3.1.2.2 VALIDEZ DEL CLÚSTER

> Cada algoritmo puede particionar datos, pero diferentes algoritmos o parámetros de entrada causan diferentes agrupaciones o revelan diferentes estructuras de agrupación. Por lo tanto, el problema de evaluar objetiva y cuantitativamente los grupos resultantes, o si la estructura de agrupamiento derivada es significativa, lo que se conoce como validación de grupos, es particularmente importante (Dubes, 1993; Gordon, 1998; Halkidi et al., 2002; Jain y Dubes, 1988). (Xu & Wunsch, 2008, pág. 221)

La validación de clúster es necesaria para determinar cuál es la eficiencia a la hora de particionar un conjunto de datos, con esto se puede corroborar una hipótesis planteada previamente. También es necesario realizar la validación previa a la aplicación de las técnicas de clustering para determinar si los datos poseen una estructura que se pueda agrupar.

Dentro de las formas de validación del clustering existen diferentes criterios a tomar en cuenta, estos son:

- **Criterio externo:** Compara la estructura de clustering obtenida y una estructura previamente especificada. Algunos índices externos son: índice rand, coeficiente Jaccard, índice Fowlkes y Mallows y estadísticas Γ.

- **Criterio interno:** Evalúa la estructura de clustering exclusivamente usando el conjunto de datos que se tiene, sin ninguna información externa.

- **Criterio relativo:** Compara una estructura de clustering determinada con otras, obtenidas de la aplicación de diferentes algoritmos de clustering o el mismo algoritmo con diferentes parámetros.

1. CRITERIO RELATIVO

Los criterios internos y externos requieren de pruebas estadísticas, lo cual puede demandar un alto rendimiento computacional. Este criterio elimina dichos requerimientos y se concentra en la comparación de los resultados de diferentes algoritmos de clustering o de uno solo con diferentes parámetros.

Con este criterio se puede resolver un problema conocido, el cual es determinar el número real de clústeres que se encuentra dentro de un conjunto de datos al cual denominaremos k. Para los algoritmos jerárquicos k nos indica donde cortar el dendrograma y para los algoritmos basados en prototipos es el parámetro más importante.

> Ya sea sobreestimación o subestimación de K afectará la calidad de los grupos resultantes. Una partición con demasiados grupos complica la verdadera estructura de agrupamiento, por lo que es difícil interpretar y analizar los resultados. Por otro lado, una partición con muy pocos grupos causa la pérdida de información y confunde la decisión final. En la siguiente sección, nos centramos en los métodos, índices y criterios utilizados para abordar este problema fundamental. (Xu & Wunsch, 2008, pág. 268)

1.1. VISUALIZACIÓN DE DATOS

Uno de los métodos más directos para estimar el valor de k es la proyección de los datos en un espacio euclidiano de dos o tres dimensiones, de esta forma una simple inspección podría proveernos de información útil sobre el número de grupos. Sin embargo, existen conjuntos de datos cuya complejidad hace que esta técnica sea insuficiente para determinar un numero óptimo de k.

1.2. INDICES DE VALIDACIÓN Y REGLAS DE DETENCIÓN

Para algoritmos que requieren de k como parámetro, una secuencia de estructuras de clustering puede ser obtenida al correr el algoritmo múltiples veces desde k_{minimo} hasta k_{maximo}.

Luego las estructuras calculadas son evaluadas por índices de validación para determinar la solución de clustering esperada eligiendo la que posea el mejor índice. Por otro lado, para los algoritmos jerárquicos los índices son conocidos como reglas de detención, ya que indican en qué nivel el dendrograma se debe cortar.

> Como estándar para evaluar grupos, estos índices combinan la información sobre la compacidad de los grupos internos y el aislamiento de los grupos externos y son funciones de ciertos factores, como el error cuadrático definido, las propiedades geométricas o estadísticas de los datos, el número de objetos de datos, la medida de disimilitud o similitud y, por supuesto, el número de grupos. (Xu & Wunsch, 2008, pág. 269)

Algunos de estos índices son:

- Índice de Caliński y Harabasz

$$CH(K) = \frac{\frac{Tr(S_B)}{k\ -1}}{\frac{Tr(S_W)}{n_E - k}}$$

$$S_B = \sum_{q=1}^{k} n_q(c_q - c_E)(c_q - c_E)^T$$

$$S_W = \sum_{q=1}^{k} \sum_{x \epsilon C_q} (x - c_q)(x - c_q)^T$$

Donde:

E: Conjunto de datos.

n_E: número de objetos de E.

$Tr(S_B)$: traza de la matriz de dispersión entre grupos.

$Tr(S_W)$: traza de la matriz de dispersión dentro del grupo.

k: número de clústeres.

C_q: Conjunto de datos en el clúster q.

c_q: Centro del clúster q.

c_E: Centro de E.

n_q: número de objetos del clúster q.

El valor de k que maximiza la ecuación $CH(K)$ sugiere una estimación de k.

- Índice Davies-Bouldin

$$R_i = \left(\frac{e_i + e_j}{D_{ij}}\right)$$

$$\bar{R} = DB(K) = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_i$$

Donde:

D_{ij}: distancia entre centroides de los clústeres i y j.

e_i, e_j : distancia promedio entre los puntos y el centroide de los clústeres i y j.

> Los datos bidimensionales aleatorios producen valores $\underline{R}$ mínimos de aproximadamente 0.6 si se prohíben los grupos de un solo miembro. Un valor de $\underline{R}$ arriba, o en el mismo rango que los mínimos obtenidos para datos distribuidos aleatoriamente, indica que una partición particular no separa los datos en grupos naturales. (Davies & W., 1979, pág. 225)

El valor de k que minimiza la ecuación $DB(K)$ indica el numero potencial de clústeres en los datos.

II.2. ANTECEDENTES DE LA INVESTIGACIÓN

II.2.1. ANTECEDENTES INTERNACIONALES

II.2.1.1. ANTECEDENTE N.º 1

El estudio "ANALYZING AND CLUSTERING NEURAL DATA" realizado en la Universidad de Boston - Massachusetts (Estados Unidos) en el año 2015 por Sinha Amit, tiene como objetivo: "… ayudar a determinar un patrón subyacente en los datos neuronales a través del clustering" (pág. 1). Para lo cual se obtuvieron los datos a través de electrodos de electrocorticografía (ECoG), después se tenía que determinar bajo que estrategia se podía analizar los datos.

> Debido a que no conocemos los mecanismos internos del comportamiento cognitivo, no tenemos una verdad básica, es decir, no hay una línea de base con la que comparar los conjuntos de datos empíricos. Por lo tanto, tratamos esto como una aplicación de aprendizaje no supervisado. (Sinha, 2015, pág. 5)

El aprendizaje no supervisado nos permite analizar datos no etiquetados y su objetivo es buscar los subgrupos que sean diferentes entre sí, pero similares entre sus registros.

> Cuando k [número de clústeres] = 3 o 4, los resultados de la agrupación de k-means continúan produciendo agrupaciones con una superposición mínima, mientras que los resultados de la agrupación espectral no son concluyentes. Para los tres casos de agrupación espectral de autoajuste, las agrupaciones resultantes parecen superponerse. Al observar los conglomerados en una representación bidimensional, existe una superposición notable entre los conglomerados en el agrupamiento espectral y el agrupamiento espectral de autoajuste en oposición a k-means. (Sinha, 2015, pág. 28)

Comentario: Cuando se plantea el análisis de cualquier tipo de datos se debe considerar las técnicas adecuadas, en este caso se eligió el aprendizaje no supervisado ya que no se pueden comparan los datos neuronales con otros. Además, existen múltiples técnicas de aprendizaje no supervisado y son capaces de mostrar diferentes resultados por lo cual es importante aplicar las técnicas de clustering y preprocesamiento de datos adecuados para los datos que se manejan.

II.2.1.2. ANTECEDENTE N.º 2

En el estudio "CLUSTERING ANALYSIS OF RESIDENTIAL LOADS" realizado en la Universidad Estatal de Kansas - Kansas (Estados Unidos) en el año 2016 por Karimi Kambiz se analiza los datos recopilados de 101 casas de Austin TX mediante clustering y se menciona que el algoritmo k-means "clasifica todas las casas en uno de los grupos midiendo su distancia cuadrada de suma al centro de cada grupo y colocándolas en el grupo con la distancia cuadrada de suma más baja" (pág. 11). Es decir, el algoritmo k-means utiliza la distancia cuadrada de suma para verificar la similitud de los registros dentro de un conjunto de datos. Además, se menciona que para este algoritmo "el número de clústeres debe estar predefinido" (pág. 11).

Previamente en la investigación Karimi Kambiz menciona que: "para elegir el número correcto de grupos, la forma más fácil es tener un rango estimado de número de grupos y realizar una prueba y error para ver cuántos grupos dan los mejores resultados" (pág. 4).

> Descubrimos que hay tres tipos de usuarios en Austin, TX, según sus patrones de uso de electricidad. Un bajo porcentaje de usuarios mantuvo sus clústeres durante todo el año, mientras que la mayoría de los usuarios cambiaron su clúster una vez. Concluimos de esto que el comportamiento del uso de electricidad no se mantiene igual, sino que cambia de una estación a otra. Este cambio puede deberse al nivel de ingresos, el uso de los sistemas fotovoltaicos, el tipo de sistemas de calefacción y refrigeración, la cantidad de diferentes aparatos eléctricos y algunos otros factores. (Karimi, 2016, pág. 26)

Comentario: El análisis de clúster realizado en esta investigación se llevó a cabo con el uso del algoritmo k-means que requiere como parámetro de entrada el conjunto de datos y el número de clústeres, este último debe ser calculado antes de aplicar el análisis final, por lo tanto, se deben de considerar técnicas que permitan elegir el numero adecuado de clústeres que existen dentro del conjunto de datos.

II.2.1.3. ANTECEDENTE N.º 3

En el estudio "CLUSTER ANALYSIS OF CHILD HOMICIDE IN SOUTH KOREA" realizado en Corea del Sur en el año 2020 por Jung KyuHee, Kim Heesong, Lee Eunsaem, Choi Inseok, Lim Hyeyoung, Lee Bongwoo, Choi Byungha, Kim Junmo, Kim Hyejeong y Hong Hyeon-Gi se aplica el análisis de clúster usando la distancia Gower a un conjunto de datos de 341 casos originales de incidentes de homicidio que involucraban a niños de 0 a 18 años del 2016 con el objetivo de "identificar la tipología del homicidio infantil en Corea del Sur" (Jung, y otros, 2020, pág. 2).

Avanzando con la investigación los investigadores también encontraron que uno de los problemas a enfrentar era el tipo de datos que se encontraron y que se debía utilizar una estrategia específica para resolverlo.

> Nuestro estudio tuvo como objetivo resolver la cuestión de los datos de heterogeneidad y derivar subgrupos significativos en los datos de homicidios infantiles de Corea del Sur. Como el conjunto de datos es mixto y contiene no solo variables continuas sino también variables binarias, ordinales y categóricas, la distancia euclidiana, que trata solo con el tipo numérico de variable, no era adecuada; por lo tanto, la distancia de Gower (Gower, 1971), diseñada para el tratamiento de datos mixtos, se calculó para medir la diferencia. Un valor bajo indica que las dos variables son similares y un valor alto indica que las dos son completamente diferentes. (Jung, y otros, 2020, pág. 6)

> Los resultados mostraron 8 perfiles diferentes dentro del conjunto de datos: tortura infantil, filicidio materno, neonaticida, muerte no relacionada con abuso previo, filicidio paterno, infanticidio paterno, infanticidio materno y asesinatos psicóticos. Dentro de estos casos previamente se habían juzgado 95 como al menos sospechosos de homicidio infantil. Además, se llega a la conclusión de que "… los perfiles derivados en este estudio pueden ser útiles en la etapa inicial de investigación y usarse como una pista para señalar la dirección de una investigación adicional" (Jung, y otros, 2020, pág. 14).

Comentario: Una evaluación de los tipos de datos es necesaria para determinar las estrategias y algoritmos que se pueden usar con el conjunto de datos, existen dos opciones: adaptar los datos a un rango numérico o utilizar una función de similitud que tenga en cuenta los valores no numéricos.

II.2.2. ANTECEDENTES NACIONALES

II.2.2.1. ANTECEDENTE Nº 1

El estudio "IMPLEMENTACIÓN DE UNA HERRAMIENTA DE ANÁLISIS DE RIESGO DE CRÉDITO BASADO EN EL MODELO DE RATING DE CRÉDITO, ALGORITMOS GENÉTICOS Y CLUSTERING JERÁRQUICO AGLOMERATIVO" realizado en Universidad Nacional Mayor de San Marcos - Lima (Perú) en el año 2017 por Ramos Martinez Henry Marcos tiene el objetivo de: "Diseñar e implementar una solución, basada en la inteligencia artificial, que genere un modelo de clasificación del riesgo de crédito de clientes comerciales de acuerdo al modelo de rating de crédito" (pág. 15), también se determinan grupos subyacentes para determinar su probabilidad de riesgo, mediante la aplicación de clustering.

En esta investigación se menciona las características de las técnicas de clustering jerárquico.

> Por otro lado, los algoritmos de clustering jerárquico se aproximan al problema de clustering a través del desarrollo de una estructura de datos basada en un árbol binario, llamada dendrograma. Una vez que el dendrograma está construido, se puede escoger automáticamente el número correcto de clústeres al dividir al árbol en diferentes niveles para obtener diferentes soluciones de clustering, sin necesidad de volver a procesar nuevamente el algoritmo de clustering. El clustering jerárquico puede ser logrado a través de dos diferentes maneras, llamadas clustering aglomerativo (o de "abajo hacia arriba") y clustering divisivo (o de "arriba hacia abajo"). (Ramos Martinez, 2017, pág. 31)

El investigador llego a la conclusión de que las técnicas de inteligencia artificial empleadas mostraron un buen resultado para generar un modelo de clasificación, también son capaces de ser interpretadas fácilmente por un experto.

Comentario: La implementación de técnicas de inteligencia artificial en un modelo de negocio son fructíferos, ya que se automatizan procesos que normalmente desarrolla el recurso humano de una entidad haciendo que estos sean más exactos y que se desarrollen en un periodo de tiempo más corto. Por otro lado, al aplicar el análisis con un algoritmo de clustering jerárquico no es necesario realizar el método de prueba y error para determinar la cantidad de patrones en el conjunto de datos.

II.2.2.2. ANTECEDENTE Nº 2

El estudio "APLICACIÓN DE LA MINERÍA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS" realizado en la Universidad Mayor de San Marcos - Lima (Perú) en el año 2015 por Mamani Rodríguez Zoraida Emperatriz tiene como objetivo "Desarrollar un prototipo que aplique minería de datos distribuida mediante el uso de un algoritmo de clustering basado en la técnica k-means" (pág. 3), donde se concluye fundamentando los beneficios que las organizaciones obtendrían con su implementación.

También se menciona un proceso que es utilizado para el análisis de datos, dentro del cual se posiciona las técnicas de clustering.

> El proceso de KDD es el proceso de usar métodos algoritmos de Minería de datos para extraer (identificar) lo que es considerado conocimiento de acuerdo a las especificaciones de medidas y umbrales, usando la base de datos junto con algún pre-procesamiento requerido, sub-muestreo y transformaciones de esa base de datos. (Mamani Rodríguez, 2015, págs. 7, 8)

Comentario: Las bases teóricas del análisis de clustering se encuentran dentro de la minería de datos y son parte del proceso KDD, juntando estos se puede analizar todo tipo de datos, y así incorporarlo en diferentes rubros o campos de especialidad, hasta el momento se han observado sus beneficios en medicina, administración, aprendizaje, entre otros.

CAPÍTULO III. METODOLOGÍA

III.1. TIPO DE INVESTIGACIÓN

El tipo de investigación a realizar en este proyecto será de tipo básica. Una investigación básica se define como aquella que, "… está dirigida a un conocimiento más completo a través de la comprensión de los aspectos fundamentales de los fenómenos, de los hechos observables o de las relaciones que establecen los entes" (CIENCIACTIVA, 2016, pág. 7).

Además, esta tendrá un enfoque cuantitativo y un nivel descriptivo que, "… únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, esto es, su objetivo no es indicar cómo se relacionan éstas" (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014, pág. 92)

En esta investigación se busca generar conocimiento más conciso sobre los casos de menores desaparecidos en el Perú mediante técnicas de aprendizaje automático no supervisado, en este caso se utilizará el clustering debido a que los datos recolectados no están etiquetados.

III.2. DISEÑO DE LA INVESTIGACIÓN

El diseño por utilizar en este proyecto de investigación está alineado con el diseño experimental, que, según Hernández Sampieri y otros: "se utilizan cuando el investigador pretende establecer el posible efecto de una causa que se manipula" (pág. 130).

En esta investigación se pretende experimentar con los datos recolectados de la página "Te Estamos Buscando" variando el número de clústeres que se utiliza como parámetro de entrada en los algoritmos de clustering de tipo particional para así comparar los resultados de los índices de validación, dicho enfoque se inspira en las técnicas de validación relativa de clustering. Esto permitirá determinar el número de patrones que existen de forma natural dentro del conjunto de datos.

III.2.1. FASE 1: RECOLECTAR DATOS

Se creo una herramienta de software usando el lenguaje de programación Python para extraer los datos abiertos de la página web www.teestamosbuscando.pe, esta herramienta realiza el proceso de web scraping que detallaremos posteriormente y obtiene a través de la lectura de los archivos HTML los datos de los perfiles de menores desaparecidos.

III.2.2. FASE 2: PRE-PROCESAMIENTO DE DATOS

Para esta fase se continuará utilizando Python y la librería Scikit-Learn para pasar los datos obtenidos por un proceso que busca integrar, limpiar, seleccionar y transformar dichos datos.

Debido a la naturaleza de los datos encontrados en la página se deben aplicar técnicas y estrategias que permitan eliminar inconsistencias como valores atípicos, datos nulos y duplicados. Además, se debe estandarizar los valores encontrados en los atributos categóricos y numéricos, para obtener un mejor análisis de clustering de los registros de datos.

III.2.2.1. LIMPIEZA DE DATOS

Dentro del conjunto de datos de menores desaparecidos encontramos atributos de tipo nominales y numéricos, por lo cual se debe adoptar una estrategia diferente para cada uno de estos dos tipos.

Primero, los datos numéricos tienen características que nos pueden ayudar en el paso de tratar los datos faltantes, para este se utiliza la media de los valores que representa un valor central dentro del rango de distribución de los datos. En cuanto a los datos ruidosos, podemos realizar un análisis visual de un diagrama de dispersión de todas las dimensiones.

Segundo, los datos nominales primero deben de ser discretizados dentro de subconjuntos con mayor relevancia; para esto se realiza un mapeo de los posibles valores iniciales que existen en el conjunto de datos y se realiza la estandarización de dichos valores. Después, los valores faltantes o nulos serán reemplazados con la moda que, al igual que la media para los datos numéricos, representa el valor central de la distribución de los datos.

Finalmente obtendremos un conjunto de datos discretizado sin valores faltantes y con una cantidad mínima de datos atípicos.

III.2.2.2. REDUCCIÓN DE DATOS

Para el paso de reducción de datos utilizaremos dos técnicas: la primera será un análisis visual de los datos de tipo numérico en diagrama de dispersión para detectar valores atípicos y así evitar la disminución de precisión en el desempeño del algoritmo de clustering y la segunda será aplicar el Análisis de Componentes Principales para reducir la dimensionalidad y eliminar cualquier ruido presente en el conjunto de datos.

III.2.2.3. TRANSFORMACIÓN DE DATOS

El algoritmo seleccionado para aplicar el análisis de clustering es k-means, este algoritmo solo trabaja con valores numéricos. Por consiguiente, todos los atributos tienen que ser transformados a valores de dicho tipo; por esta razón aplicamos un proceso de transformación que incluye dos pasos:

- **Codificación:** Consiste en dar un valor numérico a cada etiqueta que existe en atributos de tipo nominal y binario. Por lo tanto, codificaremos los valores de los atributos binarios (simétricos) a números enteros positivos, sin ningún orden en específico; finalmente la codificación de atributos nominales consistirá en representar cada posible valor con pares binarios, donde el valor 0 representa la ausencia y el 1 la existencia del atributo.

- **Normalización:** Esta estrategia de transformación de datos consiste en cambiar los valores de los atributos numéricos y encajarlos dentro de un rango predefinido, para dar el mismo nivel de significancia a todos los atributos con esto el análisis de clustering estará más homogeneizado y separará los patrones tomando en cuenta cada atributo por igual.

III.2.3. FASE 3: ANÁLISIS DE CLUSTERING Y VALIDACIÓN DE RESULTADOS

El procesamiento de los datos obtenidos de la página web consiste en aplicar algoritmos de clustering y variar los parámetros de entrada (número de clústeres), para así determinar la existencia de grupos significativos en este conjunto de datos. En este caso se utilizará el algoritmo k-means, que es uno de los más utilizados en estrategias de aprendizaje no supervisado, debido a su eficiencia con conjuntos de datos pequeños.

Por lo tanto, el flujo para procesar los datos y realizar el análisis de clustering consistirá en:

- Estimar un rango de números de clústeres

- Aplicar k-means al conjunto de datos, con los diferentes valores del rango estimado

Además, se aplicarán los índices de Caliński y Harabasz, y Davies-Bouldin a los resultados obtenidos de k-means. Para determinar el nivel de eficiencia en la segmentación de patrones del conjunto de datos de menores desaparecidos, para esto se buscará el resultado que maximice el índice Caliński y Harabasz y el que tenga un valor más cercano a 0 aplicando el índice Davies-Bouldin.

III.2.4. FASE 4: INTERPRETACIÓN DE RESULTADOS

La última fase del diseño de la investigación consistirá en mostrar las características de los patrones o subpoblaciones encontradas dentro del conjunto de datos mediante técnicas visuales.

III.3. POBLACIÓN Y MUESTRA

III.3.1. POBLACIÓN

Como población se tomará en cuenta los casos de menores desaparecidos a nivel nacional registrados en la página "Te Estamos Buscando" desde su creación, 7 de febrero del 2018, hasta el 25 de abril del 2020.

Tabla 2

Población de menores desaparecidos y encontrados.

Estado	Número de personas
Encontrados	2853
Desaparecidos	4759
Total	7612

III.3.2. MUESTRA

Teniendo en cuenta que los datos de perfiles de menores desaparecidos se encuentran en un portal web abierto, la recolección de los datos se pudo realizar sin ningún inconveniente en su totalidad con el uso de la herramienta de web scraping. Debido a la accesibilidad de los datos y al enfoque del aprendizaje automático no supervisado, el muestreo podría resultar innecesario y contraproducente ya que las técnicas de clustering se implementan a través de algoritmos matemáticos recursivos y su eficacia para agrupar los datos mejora con relación a su cantidad.

III.4. INSTRUMENTOS

Para el registro de datos no se utilizó instrumentos convencionales como cuestionarios o entrevistas, por el contrario, se elaboró una herramienta de software para recolectar los datos de manera automatizada. Esta herramienta está basada en el concepto de web scraping y fue implementada con el lenguaje de programación Python (3.7.4).

El proyecto se encuentra en un repositorio abierto de GitHub (https://github.com/royexr/te_estamos_buscando_ws).

La herramienta se conecta mediante el protocolo HTTP al portal web, para lo cual se hace uso del paquete ***Requests (https://pypi.org/project/requests/)*** que facilita las solicitudes. También se hace uso de la librería ***Beautiful Soup (https://pypi.org/project/beautifulsoup4/)*** que ayuda a navegar entre paginas HTML o archivos XML para extraer los datos de su contenido.

III.5. RECOLECCIÓN Y ANÁLISIS DE DATOS

III.5.1. TÉCNICAS DE RECOLECCIÓN DE DATOS

Para la recolección de los datos se utilizó la técnica de web scraping que consiste en emular la navegación de una persona y acceder a datos de la red mundial (World Wide Web), en específico se accederá al portal www.teestamosbuscando.com, cuyos datos son abiertos para el público en general.

III.5.1.1. ESTRUCTURA DE DATOS

Los perfiles de menores desaparecidos son registrados para crear nota de alertas en la página web, dichos perfiles cuentan con múltiples datos como se puede apreciar en la siguiente imagen.

Figura 7

Ejemplo de perfil de menor desaparecido.

Nota. Fuente: (Ministerio del Interior, 2021)

Para extraer estos datos, se creó una clase en Python con las siguientes características:

Tabla 3

Atributos de perfil de menor desaparecido.

Número	Atributo	Tipo	Descripción
1	Edad	Número	Edad del menor desaparecido
2	Circunstancias	Cadena	Circunstancias en las que el menor desapareció
3	Vestimenta	Cadena	Vestimenta con la que el menor desapareció
4	Nombre	Cadena	Nombre completo del menor
5	Genero	Cadena	Genero del menor
6	Nombre del informante	Cadena	Nombre de la persona que realizo la denuncia
7	Teléfono del informante	Número	Teléfono de la persona que realizo la denuncia
8	Departamento	Cadena	Departamento donde se registró la denuncia
9	Provincia	Cadena	Provincia donde se registró la denuncia
10	Distrito	Cadena	Distrito donde se registró la denuncia
11	Cabello	Cadena	Color de cabello del menor
12	Boca	Cadena	Forma o tamaño de la boca del menor
13	Ojos	Cadena	Color de ojos del menor
14	Nariz	Cadena	Forma de la nariz del menor
15	Raza	Cadena	Raza del menor
16	Estatura	Cadena	Estatura del menor
17	Fecha de reporte	Cadena	Fecha de registro de la denuncia
18	Fecha de desaparición	Cadena	Fecha en la que desapareció el o la menor
19	Url	Cadena	Url de la nota de alerta

Para extraer cada atributo de la clase se utilizó cadenas de búsqueda del API web, por lo cual se tuvo que crear un archivo de configuración con cada atributo seleccionado a recolectar.

III.5.1.2. FUNCIONAMIENTO DE LA HERRAMIENTA

Después de haber registrado las cadenas de búsqueda para cada atributo del perfil del menor, se procede a crear algoritmos para iterar entre páginas, explorar toda la página, extraer y guardar los datos.

III.5.1.2.1. FUNCIÓN INICIAL

En esta función se hace referencia al identificador de la página registrada dentro del archivo de configuración en formato YAML para pasarla a la función scraper como parámetro.

Figura 8

Función inicial

```python
if __name__ == '__main__':
    # Declarar id del sitio para extraer los query selectors
    site_uid = 'teestamosbuscando'

    #Ejecutar scraper
    _scraper(site_uid)
```

III.5.1.2.2. FUNCIÓN SCRAPER

Esta función se encarga de extraer las cadenas de búsqueda del API web según al identificador de la página, para luego iterar dentro de cada página (según al rango establecido) y obtener los enlaces para cada perfil. Este proyecto hace uso de paradigma de programación orientada a objetos para plasmar como objetos la página inicial (que contiene los resúmenes de los perfiles y esta enumerada) y la página de la persona (donde se detalla los datos registrados del menor desaparecido).

Seguidamente, la función llama a otra que se encarga de obtener los datos del perfil del menor desaparecido, lo almacena de forma temporal en un arreglo y finalmente llama a la función que se encargará de almacenar los datos.

Figura 9

Función Scraper

```python
def _scraper(site_uid):
    # Obtener raiz de url
    host = config()['sites'][site_uid]['url']

    # Crear variable temporal para almacenar perfiles
    people_profiles = []

    # Iterar desde 0 hasta la cantidad de paginas deseadas
    for x in range(0, 298):
        # Crear nodo de url
        node = '/desaparecidos?page=' + str(x)

        logging.info('Beginning scraper for {}'.format(host + node))

        # Crear objeto de pagina inicial
        homepage = missing_people.HomePage(site_uid, host + node)

        # Iterar por cada enlace de perfil encontrado
        for link in homepage.people_links:
            # Extraer datos de perfil
            person_profile = _fetch_person_profile(site_uid, host, link)

            if person_profile:
                logger.info('Person profile fetched!!')
                people_profiles.append(person_profile)

    # Guardar datos de perfil
    _save_people_profiles(site_uid, people_profiles)
```

III.5.1.2.3. FUNCIÓN OBTENER PERFIL

Esta función se encarga de obtener los datos de los perfiles de menores desaparecidos, para lo cual utiliza el objeto que abstrae la estructura de datos de la persona. Finalmente verifica si el perfil contiene el campo de nombre para retornarlo o devolver un valor nulo en caso contrario.

Figura 10

Función obtener perfil

```python
def _fetch_person_profile(site_uid, host, link):
    logger.info('Start fetching article at {}'.format(link))

    # Inicializar variable contenedora del perfil
    person_profile = None
    try:
        # Crear objeto con datos de la persona desaparecida
        person_profile = missing_people.PersonPage(site_uid, _build_link(host, link))
    except (HTTPError, MaxRetryError) as e:
        logger.warning('Error while fetching the person info', exc_info=False)

    # Si la persona no tiene datos de nombre, no se registra
    if person_profile and not person_profile.name:
        logger.warning('Invalid person profile. There is no name')
        return None

    return person_profile
```

III.5.1.2.4. FUNCIÓN GUARDAR PERFIL

Esta función recibe como parámetros el arreglo que contiene los perfiles recolectados y el id de la página, con estos datos se crea el archivo final a entregar formateando el nombre del archivo según la fecha de ejecución del programa. El archivo final se devuelve en formato CSV con cabeceras de los atributos del perfil del menor desaparecido.

Figura 11

Función guardar perfil

```
def _save_people_profiles(site_uid, people_profiles):
    # Declarar variable de fecha de ejecucion
    now = datetime.datetime.now().strftime('%Y_%m_%d')
    # Declarar variable de nombre de archivo
    out_file_name = '{site_uid}_{datetime}_people_profiles.csv'.format(site_uid=site_uid,
datetime=now)

    # Crear cabeceras de archivo CSV
    csv_headers = list(filter(lambda property: not property.startswith('_'),
dir(people_profiles[0])))

    # Abrir archivo en modo de escritura
    with open(out_file_name, mode='w+') as f:
        # Crear escritor para archivos CSV
        writer = csv.writer(f)
        # Escribir fila con cabeceras
        writer.writerow(csv_headers)

        #Iterar entre los perfiles obtenidos
        for person_profile in people_profiles:
            #Declarar fila con datos del perfil
            row = [str(getattr(person_profile, prop)) for prop in csv_headers]
            # Escribir fila
            writer.writerow(row)
```

III.5.2. TÉCNICAS DE ANÁLISIS DE DATOS

Los datos obtenidos serán analizados mediante técnicas de clustering y los resultados serán analizados con los índices de validación (índice Caliński y Harabasz e índice Davies-Bouldin). Utilizaremos gráficos y matrices para mostrar la distribución de los datos y los resultados de la validación.

Previo al análisis se debe realizar pasos de preprocesamiento a los datos recolectados inicialmente, como se menciona en el Capítulo I, para obtener conocimiento. Por lo tanto, para mejorar la calidad del conocimiento producido primero realizaremos los siguientes pasos:

III.5.2.1. INTEGRACIÓN DE DATOS

Consiste en combinar múltiples fuentes datos, para el caso de estudio se suscitan dos conjuntos de datos: menores desaparecidos y menores encontrados.

III.5.2.2. LIMPIEZA DE DATOS

Este paso consiste en remover ruido y datos inconsistentes, observando el conjunto de datos recolectados podemos mostrar que existe una cierta cantidad de datos faltantes por atributo de los perfiles que se muestran en la siguiente tabla.

Tabla 4

Número de datos faltantes por atributo.

Número	Atributo	Número de datos faltantes
1	Genero	38
2	Edad	82
3	Cabello	191
4	Boca	358
5	Ojos	113
6	Nariz	379
7	Estatura	275
8	Raza	1272

Además, también existen inconsistencias tipográficas en los datos recolectados, por lo cual nos encontramos con valores que representan un mismo tipo de raza, pero con diferentes nombres asociados, por ejemplo: MEZTIZA, MESTIZA, MESTIZO. Para agrupar adecuadamente estos valores, utilizaremos expresiones regulares que buscaran términos similares y cambiaran su valor con uno por defecto.

III.5.2.3. SELECCIÓN DE DATOS

Este paso consiste en seleccionar los atributos del conjunto de datos que nos serán relevantes para la producción de conocimientos. Dentro de nuestro conjunto de datos se seleccionaron ocho atributos del perfil los cuales son: edad, estatura, color de ojos, genero, color de cabello, nariz, raza.

Al aplicar los métodos de análisis podremos determinar los patrones (grupos, clústeres) de menores desaparecidos según sus características físicas.

III.5.2.4. TRANSFORMACIÓN DE DATOS

Este paso consiste en transformar los valores de los perfiles recolectados, para lo cual utilizaremos múltiples técnicas.

1) **Codificación:** los datos categóricos se deben de transformar en valores numéricos. Por ejemplo: el género de los menores varía entre "femenino" y "masculino", por lo tanto, se codificarán como 0 y 1.

2) **Normalización o Escalamiento:** los datos ahora son todos numéricos por lo cual es necesario normalizarlos dentro de un rango de valores, para que ningún atributo sea más relevante que otro dentro del análisis de clustering.

III.5.2.5. PROCESAMIENTO O ANÁLISIS

En este paso aplicaremos el análisis de datos mediante clustering, que busca identificar los grupos de datos subyacentes de un conjunto de datos.

CAPÍTULO IV. RESULTADOS

El objetivo del presente estudio fue elaborar una investigación descriptiva que nos permita determinar los clústeres (patrones) dentro del conjunto de datos de perfiles de menores desaparecidos en el Perú mediante técnicas de aprendizaje no supervisado, para esto se elaboró una secuencia de 4 etapas detalladas en el punto III.2.

IV.1. ETAPA 1: RECOLECTAR DATOS

Para aplicar los métodos de aprendizaje no supervisado se requiere de una base de datos cuyos registros sean significativos para el fin de la investigación, en este caso se extrajo todos los datos encontrados en la página web "Te Estamos Buscando" (www.teestamosbuscando.com), tanto de perfiles de desaparecidos como de encontrados.

Al aplicar la herramienta de web scraping, detallada en el párrafo III.5.1. Se obtuvo un conjunto de datos con un total de 4759 registros con 19 atributos (4759 x 19) de menores desaparecidos y 2853 registros con 19 atributos (2853 x 19) de menores encontrados. Los atributos recolectados de los perfiles de menores desaparecidos se pueden observar en la **Tabla 3**.

IV.2. ETAPA 2: PRE-PROCESAMIENTO DE DATOS

Los datos recolectados inicialmente contienen diferentes inconsistencias, por lo cual se debe seguir los pasos detallados en el punto III.5.2, para asegurar que el resultado del análisis sea el más optimo posible.

IV.2.1. INTEGRACIÓN DE DATOS

Debido a que se recolectaron datos de dos diferentes orígenes (menores desaparecidos y menores encontrados), estos dos conjuntos se deben integrar en uno solo para mejorar la precisión del modelo de clustering.

Por lo tanto, se unifico los dos conjuntos de datos cuyo resultado fue un conjunto de datos con 7612 registros y 19 atributos (7612 x 19).

IV.2.2. LIMPIEZA DE DATOS

IV.2.2.1. ELIMINAR DUPLICADOS

El primer paso en la limpieza de datos es descartar los datos repetidos o duplicados, en este caso podemos filtrar los perfiles por nombres para determinar si hay duplicidad de registros, en la **Tabla 5** podemos observar algunos de los perfiles duplicados agrupados por nombre.

Tabla 5

Perfiles duplicados por nombre.

Nombre	Cantidad
Lizeth Rodríguez Ataucusi	4
Dayana Michelle Gallo Postigo	4
Janet Andrea Bautista Mamani	3
Alex Wilfredo Farfán Soto	3
…	…

Al borrar los registros duplicados la cantidad de registros fue reducida hasta 7006.

IV.2.2.2. FORMATEAR DATOS CATEGORICOS Y NUMÉRICOS

Luego extraemos la descripción de los datos integrados, obteniendo así la **Tabla 6** que contiene los tipos de datos de cada atributo registrado en el perfil del menor.

Tabla 6

Tipos de dato por atributo del conjunto de datos inicial.

Atributo	Tipo de dato
Edad	Número
Circunstancias	Cadena
Vestimenta	Cadena
Fecha de denuncia	Cadena
Fecha de desaparición	Cadena
Departamento	Cadena
Provincia	Cadena
Distrito	Cadena
Ojos	Cadena
Género	Cadena
Cabello	Cadena
Estatura	Cadena
Nombre del informante	Cadena
Teléfono del informante	Cadena
Boca	Cadena
Nombre	Cadena
Nariz	Cadena
Raza	Cadena
Url	Cadena

La limpieza de datos comenzó con la eliminación de datos inconsistentes, se mostrarán a continuación los cambios hechos por cada atributo que poseía inconsistencias.

IV.2.2.2.1. EDAD Y ESTATURA

Dentro de la descripción inicial del conjunto de datos, podemos observar que el campo de "Estatura" se clasifica como cadena y no como número. Dentro de los valores inconsistentes podemos encontrar: Baja, Mediana, ALTA, 1.50 APROX., 1.40 MTS, 1 METRO, etc.

El primer paso consistirá en extraer solo los números de todos estos valores y estandarizarlos en centímetros. Al aplicar el cambio podremos extraer una descripción numérica del conjunto de datos plasmados en la **Tabla 7**.

Tabla 7

Descripción de atributos numéricos (inicial).

	Edad	Estatura (cm)
Recuento	6928	6731
Media	13.94	151.20
Desviación estándar	2.90	16.29
Valor mínimo	0	32
Valor máximo	38	193

Con esta descripción de los atributos numéricos podemos ver en la **Figura 12** que existen datos en el campo de edad que rebasan el límite máximo para considerar a una persona menor de edad. Por lo tanto, realizamos un filtrado simple para eliminar los datos atípicos del atributo y da como resultado el contenido de la **Tabla 8** y los registros se ubican en la **Figura 13**.

Figura 12

Diagrama de dispersión (edad x altura) (inicial).

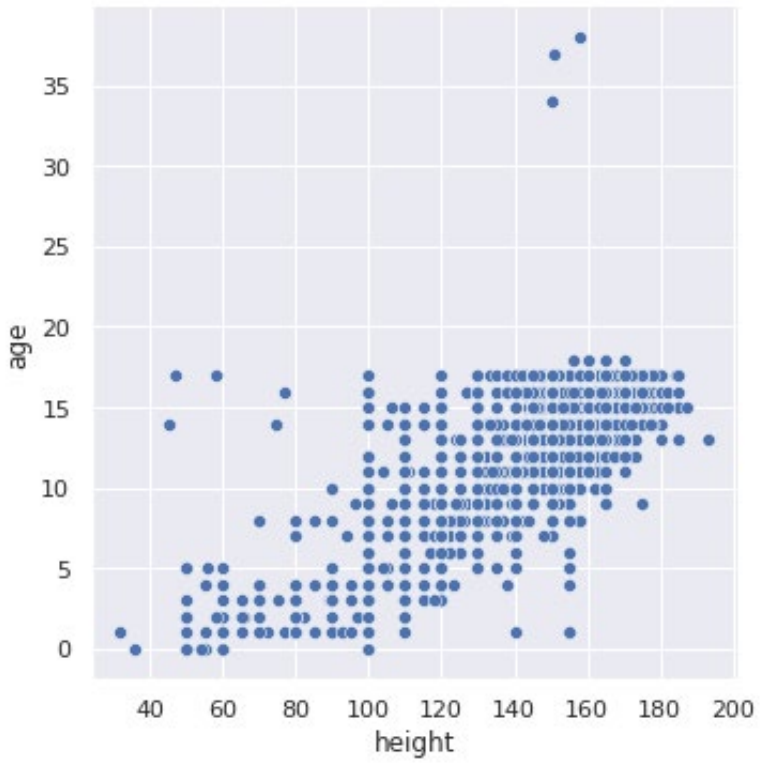


Figura 13

Diagrama de dispersión (edad x altura) (sin valores atípicos).

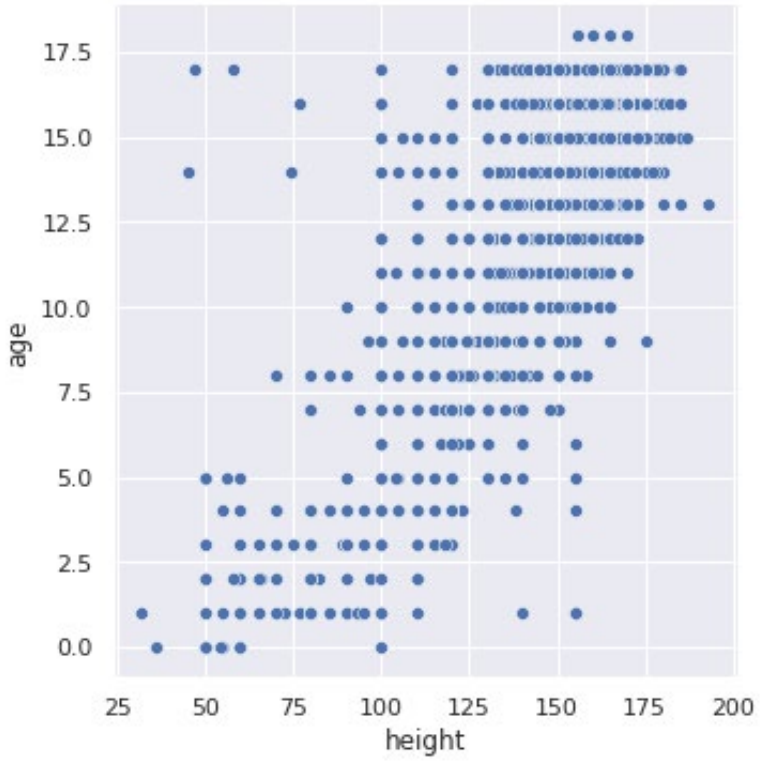


Tabla 8

Descripción de valores numéricos (sin valores atípicos).

	Edad	Estatura (cm)
Recuento	6925	6925
Media	13.82	150.82
Desviación estándar	2.83	16.19
Valor mínimo	0	32
Valor máximo	18	193

IV.2.2.2.2. OJOS

Al extraer los valores del atributo “ojos” del conjunto de datos, se obtuvo los datos de la **Tabla 9**.

Tabla 9

Valores del atributo "ojos" (inicial).

Valor	Cantidad
COLOR NEGRO	1798
NEGROS	1623
PARDOS	642
COLOR NEGRO	530
…	…
CHICO CLAROS	1
MARRON NEGRO ACHINADOS	1

Observando las diferentes variaciones de valores del atributo se decidió agrupar los valores entre un rango más adecuado, el cual contenía los colores: negro, pardo, ámbar, avellana, verde, azul y gris. Además, los valores inconsistentes como, por ejemplo: “Pardo y negro”, se transformaron en valores nulos. Obteniendo así los registros de la **Tabla 10**.

Tabla 10

Valores del atributo "ojos" (formateados).

Valor	Cantidad
NEGRO	4294
PARDO	1910
GRIS	103
AMBAR	84
AVELLANA	25
VERDE	22
AZUL	2
NULO	485

IV.2.2.2.3. CABELLO

Al extraer los valores del atributo “cabello” del conjunto de datos, se obtuvo los datos de la **Tabla 11**.

Tabla 11

Valores del atributo "cabello" (inicial).

Valor	Cantidad
COLOR NEGRO	2283
LACIO NEGRO	492
NEGRO	381
NEGROS	352
…	…
Recortado por los costados	1
CASTAÑO PINTADO DE RUBIO LACIOS	1

Podemos observar que existen valores que podrían ser agrupados en valores más significativos. Por ejemplo: “COLOR NEGRO” y “LACIO NEGRO”, para esta agrupación se tomó como parámetro principal el color que contenía cada valor creando así un rango de valores que contenía los tipos de cabello: negro, marrón, rubio, rojo, azul, gris y negro. Obteniendo así los registros de la **Tabla 12**.

Tabla 12

Valores del atributo "cabello” (formateados).

Valor	Cantidad
NEGRO	4861
MARRON	868
RUBIO	60
ROJO	29
GRIS	10
AZUL	2
NULO	1095

IV.2.2.2.4. BOCA

Al extraer los valores del atributo “boca” del conjunto de datos, se obtuvo los datos de la **Tabla 13**.

Tabla 13

Valores del atributo "boca" (inicial).

Valor	Cantidad
MEDIANA	3863
PEQUEÑA	880
NORMAL	522
GRANDE	400
…	…
MEDIANA LABIOS CARNOSOS	1
Mediana con un lunar	1

Observando las diferentes variaciones de valores del atributo se decidió agrupar los valores entre un rango más adecuado, el cual varía entre: pequeña, mediana y grande, ya que el valor más predominante es el tamaño. Obteniendo así los registros de la **Tabla 14**.

Tabla 14

Valores del atributo "boca" (formateados).

Valor	Cantidad
MEDIANA	4255
PEQUEÑA	1568
GRANDE	766
NULO	336

IV.2.2.2.5. NARIZ

Al extraer los valores del atributo "nariz" del conjunto de datos, se obtuvo los datos de la **Tabla 15**.

Tabla 15

Valores del atributo "nariz" (inicial).

Valor	Cantidad
RECTA	3457
AGUILEÑA	645
NORMAL	610
ANGULAR	335
…	…
CHICA (ÑATA)	1
Perfilado	1

Podemos observar que los valores no son consistentes, en este caso se optó por agrupar los valores en: pequeña, mediana y grande. Obteniendo así los registros de **Tabla 16**.

Tabla 16

Valores del atributo "nariz" (formateados).

Valor	Cantidad
MEDIANA	4896
GRANDE	859
PEQUEÑA	809
NULO	361

IV.2.2.2.6. RAZA

Al extraer los valores del atributo "raza" del conjunto de datos, se obtuvo los datos de la **Tabla 17**.

Tabla 17

Valores del atributo "raza" (inicial).

Valor	Cantidad
MESTIZA	4536
BLANCA	802
TRIGUEÑA	574
TRIGEÑA	107
…	…
Caucasico	1
MULATA	1

Observando las diferentes variaciones de valores del atributo se decidió agrupar los valores entre un rango más adecuado, el cual varía entre: blanca, mestiza, negra y trigueña. Obteniendo así los registros de la **Tabla 18**.

Tabla 18

Valores del atributo "raza" (formateados).

Valor	Cantidad
MESTIZA	4197
BLANCA	771
TRIGUEÑA	767
NEGRA	22
NULO	1168

IV.2.2.2.7. GÉNERO

Dentro de los valores encontrados en el atributo género no se encontraron inconsistencias, pero si una cantidad importante de datos nulos como se observa en la **Tabla 19**.

Tabla 19

Valores del atributo "género" (formateados).

Valor	Cantidad
FEMENINO	5037
MASCULINO	1857
NULO	31

Finalmente obtenemos las estadísticas descriptivas de cada atributo por tipo como se observa en la **Tabla 20** y la **Tabla 21**.

Tabla 20

Estadísticas descriptivas de atributos binarios (con valores nulos).

	Género
Recuento	6894
Valores únicos	2
Moda	Femenino
Frecuencia (Moda)	5037

Tabla 21

Estadísticas descriptivas de atributos nominales (con valores nulos).

	ojos	cabello	boca	nariz	raza
Recuento	6440	5830	6589	6564	5757
Valores únicos	7	6	3	3	4
Moda	NEGRO	NEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia (Moda)	4294	4861	4255	4896	4197

IV.2.2.3. LLENANDO VALORES FALTANTES

El siguiente paso en la limpieza de datos es llenar los datos faltantes. Para esto utilizaremos dos estrategias diferentes de acuerdo con el tipo de atributo que se maneje, en este caso tenemos: atributos nominales, binarios y numéricos.

Para llenar los valores faltantes de los atributos nominales y binarios utilizaremos la estrategia de replicar el valor más frecuente y para los valores numéricos utilizaremos la estrategia de replicar las medias como se indica en el punto II.1.1.1.1.1. La clasificación de los atributos por tipo de datos es:

- **Numéricos:** edad y altura.
- **Nominales:** cabello, ojos, raza, boca y nariz.
- **Binarios (simétrico):** género.
- **Únicos:** circunstancias, vestimenta, fecha de denuncia, fecha de desaparición, departamento, provincia, distrito, nombre del informante, teléfono del informante, nombre del desaparecido y url.

Después de aplicar las técnicas para rellenar los valores faltantes obtenemos los resultados de la **Tabla 22**, la **Tabla 23** y la **Tabla 24**.

Tabla 22

Estadísticas descriptivas de atributos numéricos (sin valores nulos).

	Edad	**Altura**
Recuento	6925	6925
Media	13.81	150.81
Desviación estándar	2.84	16.11
Valor mínimo	0	50
Valor máximo	8	193

Tabla 23

Estadísticas descriptivas de atributos binarios (sin valores nulos).

	Género
Recuento	6925
Valores únicos	2
Moda	Femenino
Frecuencia (Moda)	5068

Tabla 24

Estadísticas descriptivas de atributos nominales (sin valores nulos).

	ojos	cabello	boca	nariz	raza
Recuento	6925	6925	6925	6925	6925
Valores únicos	7	6	3	3	4
Moda	NEGRO	NEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia (Moda)	4779	5956	4591	5257	5365

IV.2.3. TRANSFORMACIÓN DE DATOS

Debido a que los datos procesados pasaran por el algoritmo k-means es necesario asignar una representación numérica para cada atributo, la estrategia utilizada para este paso se detalla en el punto III.2.2.3.

El resultado de los pasos de preprocesamiento previos nos permite mostrar todos los atributos como valores numéricos como se muestra en la **Tabla 25**.

Tabla 25

Estadísticas descriptivas del conjunto de datos (después del preprocesamiento).

	Edad	Estatura	Género	0	1	2	…	150
Recuento	6925	6925	6925	6925	6925	6925	…	6925
Media	0.7738	0.7836	0.2682	0.0121	0.0036	0.0003	…	0.0038
Desviación estándar	0.1590	0.0822	0.4430	0.1095	0.0600	0.0170	…	0.0612
Valor mínimo	0	0	0	0	0	0	…	0
Valor máximo	1	1	1	1	1	1	…	1

IV.2.4. REDUCCIÓN DE DATOS

Debido a la naturaleza del análisis de clustering debemos utilizar atributos no únicos o cuyo rango de variación entre valores no sea muy amplio, estos son: atributos numéricos (altura y edad), binarios (género) y nominales (boca, cabello, nariz, ojos, y raza). Por lo tanto, para extraer patrones con mayor precisión se tienen que descartar los atributos únicos y aplicar el algoritmo de clustering a los atributos numéricos, nominales y binarios.

Además, se utilizo el algoritmo de Análisis de Componentes Principales (PCA) para reducir las dimensiones del conjunto de datos cuyo resultado se muestra en la **Tabla 26**.

Tabla 26

Estadísticas descriptivas del conjunto de datos (redimensionado).

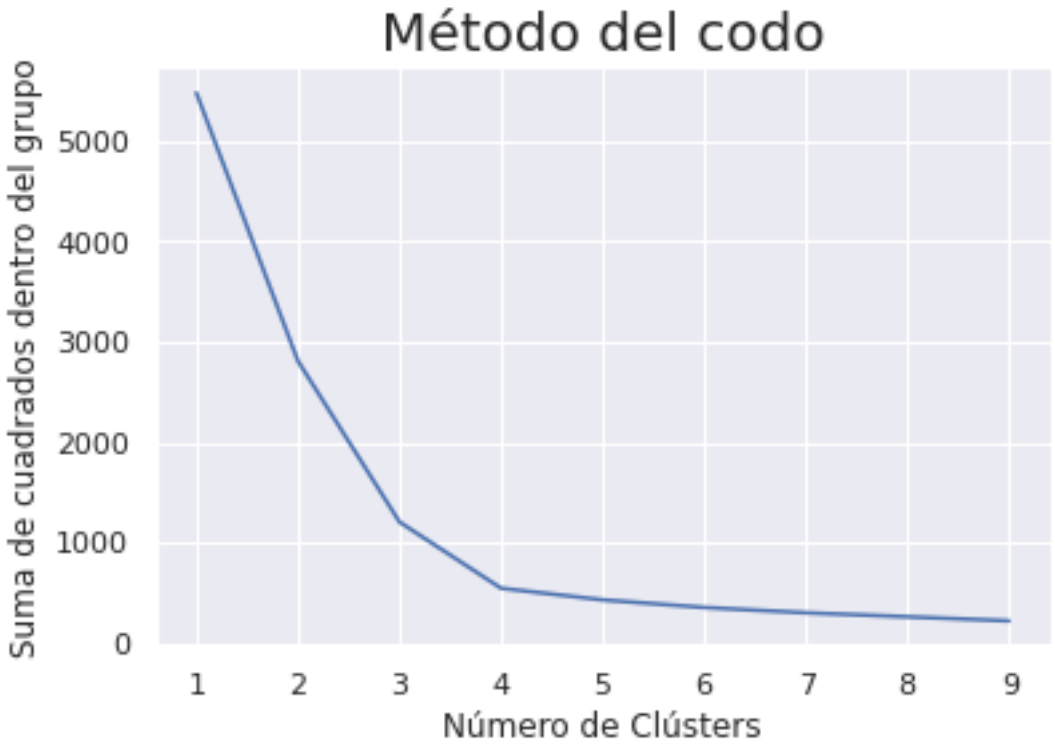
	PC-1	PC-2
Recuento	6.925000e+03	6.925000e+03
Media	-1.008900e-16	6.305425e-17
Desviación estándar	6.492898e-01	6.189703e-01
Valor mínimo	-7.284487e-01	-7.583345e-01
Valor máximo	1.508745e+00	1.763965e+00

IV.3. ETAPA 3: ANÁLISIS DE CLUSTERING Y VALIDACIÓN DE RESULTADOS

Primero, para determinar un rango de número de clústeres (k) y realizar la validación con los índices seleccionados se utilizará la técnica visual definida como método de codo (elbow method). Esta técnica consiste en utilizar la suma de cuadrados de las distancias de los registros hacia el centro de su clúster variando el número de clústeres, con la cual se proyecta una curva en un plano bidimensional y los valores que pueden ser considerados como el número de clústeres adecuado para el algoritmo se encuentran en la parte central de la curva.

Figura 14

Método de codo aplicado al conjunto de datos de menores desaparecidos.



Dentro del diagrama plasmado en la **Figura 14** podemos observar que los puntos centrales de la curva varían entre 3 a 6 clústeres, con este rango de valores aplicamos el algoritmo k-means al conjunto de datos y luego utilizamos los índices de validación (Caliński-Harabasz y Davies-Bouldin), así obtenemos:

Tabla 27

Resultados de índices de validación.

Número de clústeres (k)	Índice de Caliński y Harabasz	Índice de Davies-Bouldin
3	11128.800	0.56246294
4	16743.733	0.47875169
5	16614.093	0.62008086
6	15631.994	0.71642489

Según los índices de validación el número adecuado de clústeres que se elegirá para aplicar k-means es aquel que maximice el índice de Caliński y Harabasz y el que sea más próximo a 0 según el índice de Davies-Bouldin. Por lo tanto, como se observa en la **Tabla 27** el número de clústeres que cumple con las condiciones de ambos índices de validación es 4.

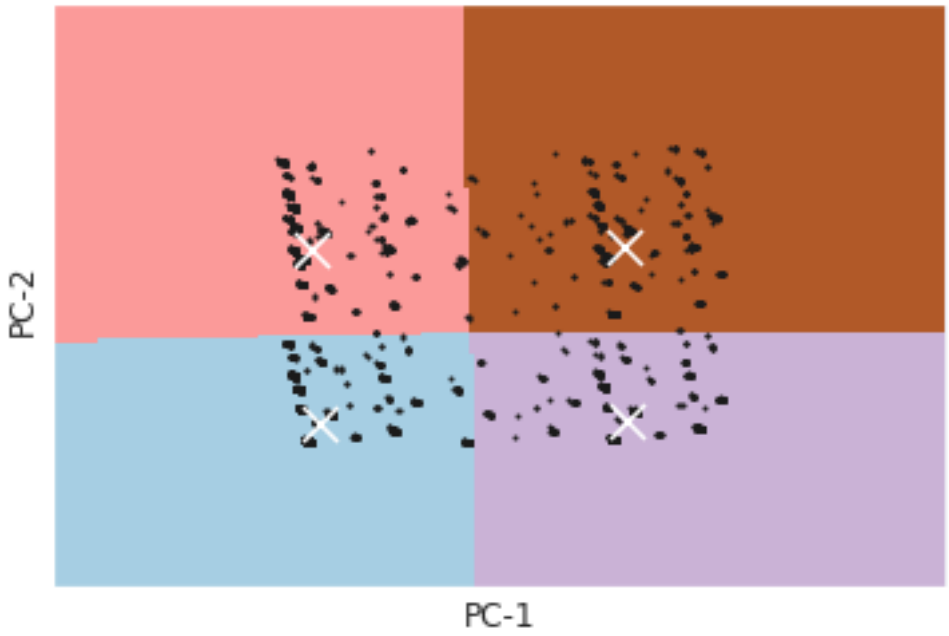
IV.4. FASE 4: INTERPRETACIÓN DE RESULTADOS

IV.4.1. DISTRIBUCIÓN DE CLÚSTERES

Primero se muestra mediante un diagrama dispersión del conjunto de datos la distribución de los clústeres o su separación en un plano bidimensional, para lo cual se utiliza el conjunto de datos redimensionado o reducido usando el algoritmo de PCA. Podemos observar en la **Figura 15** y la **Figura 16** la distribución de los registros según los componentes principales (PC-1, PC-2).

Figura 15

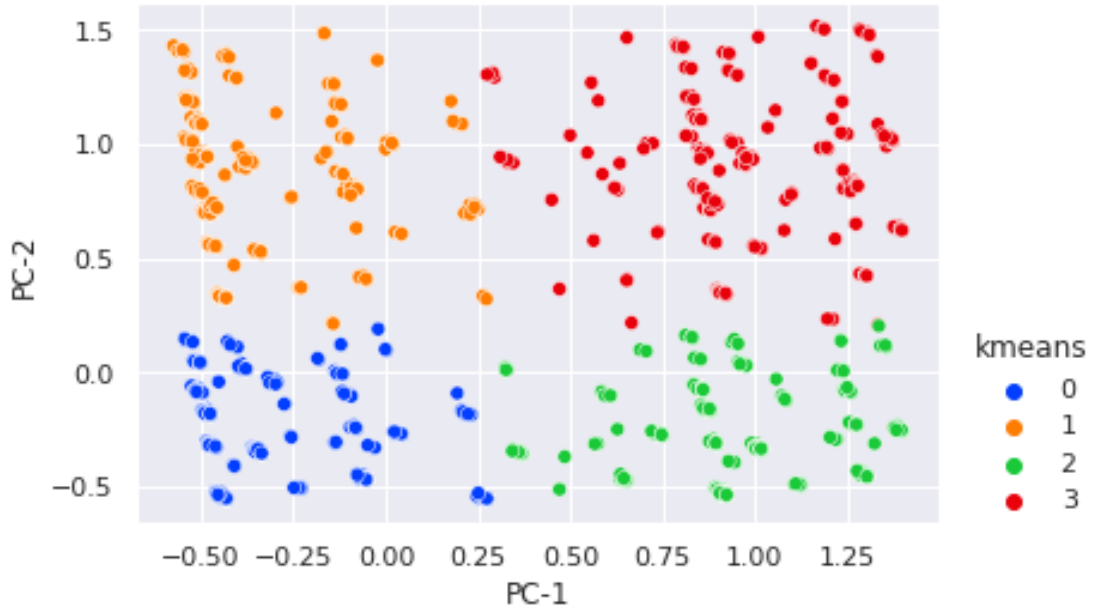
K-means aplicado a los datos de menores desaparecidos (Datos reducidos con PCA).



Nota. Los centroides están marcados con una cruz blanca.

Figura 16

Diagrama de dispersión (Componente principal 1 x Componente principal 2).



IV.4.1.1. CARACTERÍSTICAS DE LOS CLUSTERES

Después de la distribución de los registros entre cuatro clústeres, viene el paso de describir cada clúster encontrado en el conjunto de datos. Por consiguiente, exponemos los datos de cada clúster en las tablas más adelante.

La distribución de los datos en clústeres nos permite analizar visualmente sus diferencias entre las dimensiones (atributos). Por lo tanto, utilizaremos el grafico de caja y bigotes, que nos muestra los cuartiles y la amplitud de los valores con respecto a la edad, y el grafico de barras, que mostrara la distribución de cantidades de atributo por clúster.

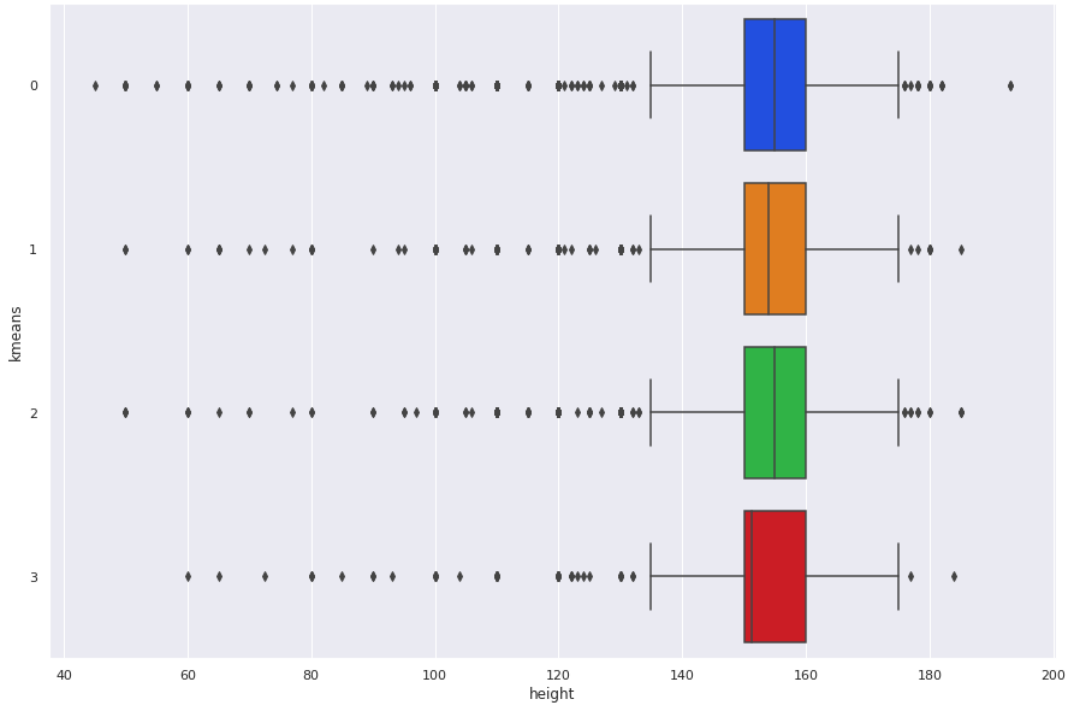
IV.4.1.2. ATRIBUTOS NUMÉRICOS

IV.4.1.2.1. ALTURA

Podemos observar en la **Figura 17** que los promedios de altura entre los clústeres difieren un poco, teniendo el promedio mas bajo en el clúster 4 (rojo).

Figura 17

Diagrama de cajas (altura x clúster).

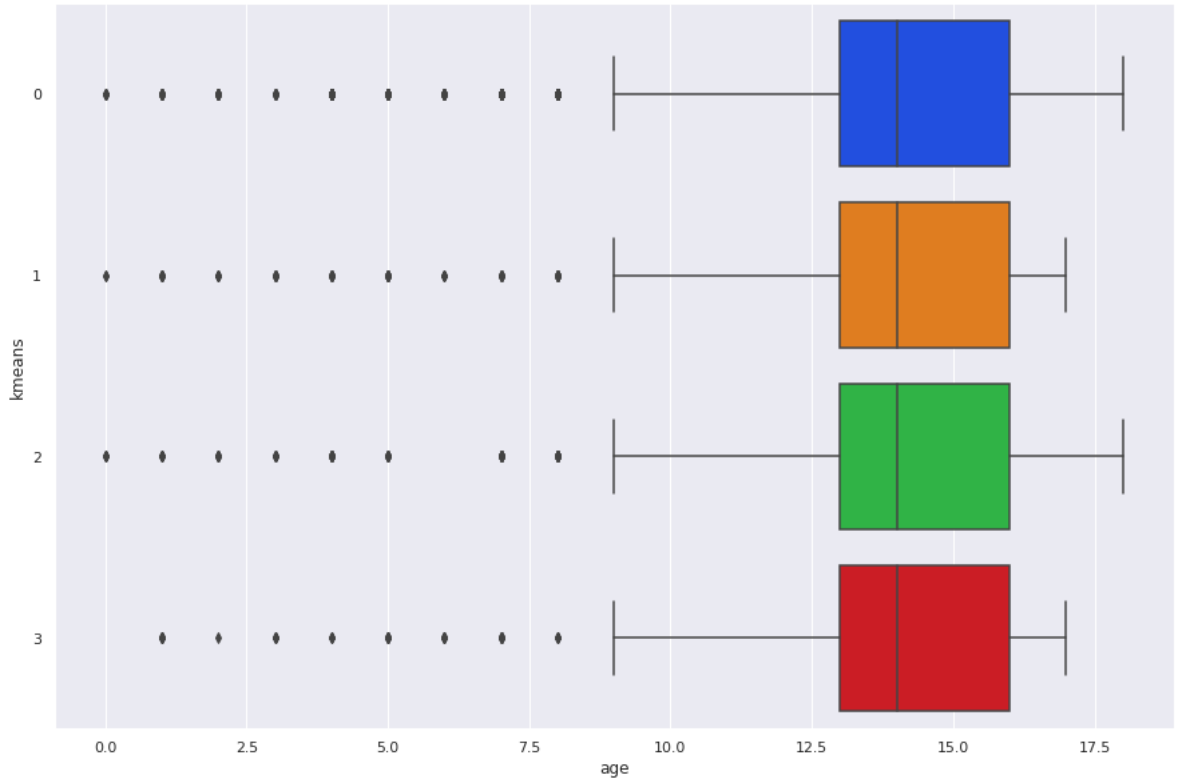


IV.4.1.2.2. EDAD

En la **Figura 18** por otro lado, podemos observar que las medias de los valores de edad no difieren lo suficiente como para observar detalles en el grafico.

Figura 18

Diagrama de cajas (edad x clúster).



IV.4.1.3. ATRIBUTOS NOMINALES

IV.4.1.3.1. COLOR DE OJOS

Dentro de los clústeres de la **Tabla 28** se observa que el atributo color de ojos, tiene una mayoría numérica para el valor "NEGRO" que representan casi el 68% de todos los registros y la minoría está en los valores "AVELLANA" y "VERDE" con 0.26% y 0.20% respectivamente. Además, podemos observar que el Clúster 3 tiene el promedio de edad mayor y el Clúster 4 tiene el promedio de edad menor.

Tabla 28

Resumen de distribución (Color de ojos x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
AMBAR	53	0.77%	13.91	17	0.25%	13.47	22	0.32%	14.77	8	0.12%	10
AVELLANA	11	0.16%	14.09	3	0.04%	14.67	2	0.03%	14.5	2	0.03%	13.5
AZUL	1	0.01%	14	-	0%	-	1	0.01%	14	-	0%	-
GRIS	59	0.85%	13.92	22	0.32%	14.14	29	0.42%	14.93	13	0.19%	14
NEGRO	2210	31.91%	13.75	1134	16.38%	13.82	868	12.53%	13.75	495	7.15%	13.82
PARDO	921	13.30%	13.94	466	6.73%	14.02	396	5.72%	13.93	178	2.57%	14.12
VERDE	4	0.06%	15.25	5	0.07%	14.4	3	0.04%	13.33	2	0.03%	16
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 19

Diagrama de cajas (Color de ojos x Edad).

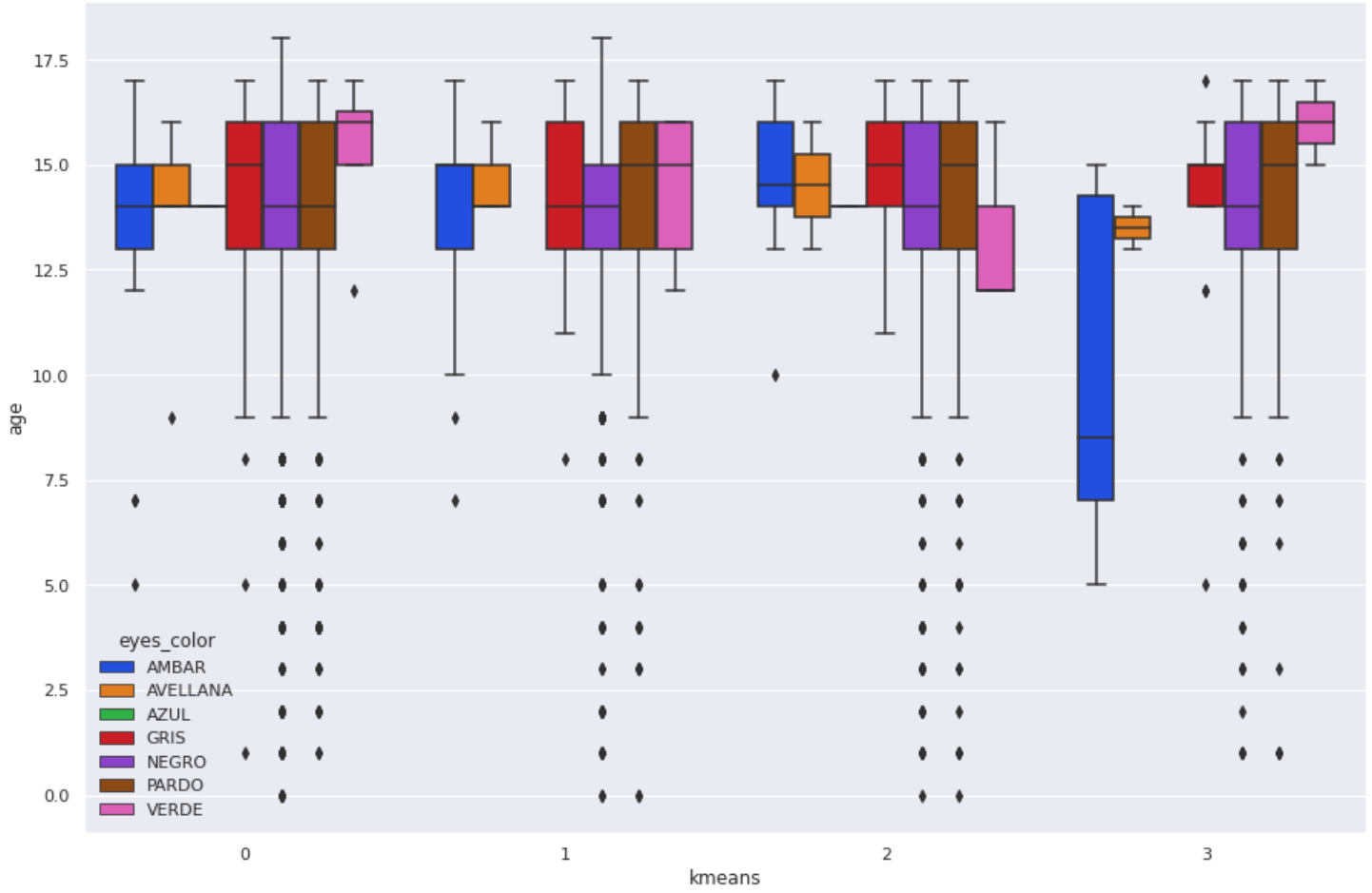
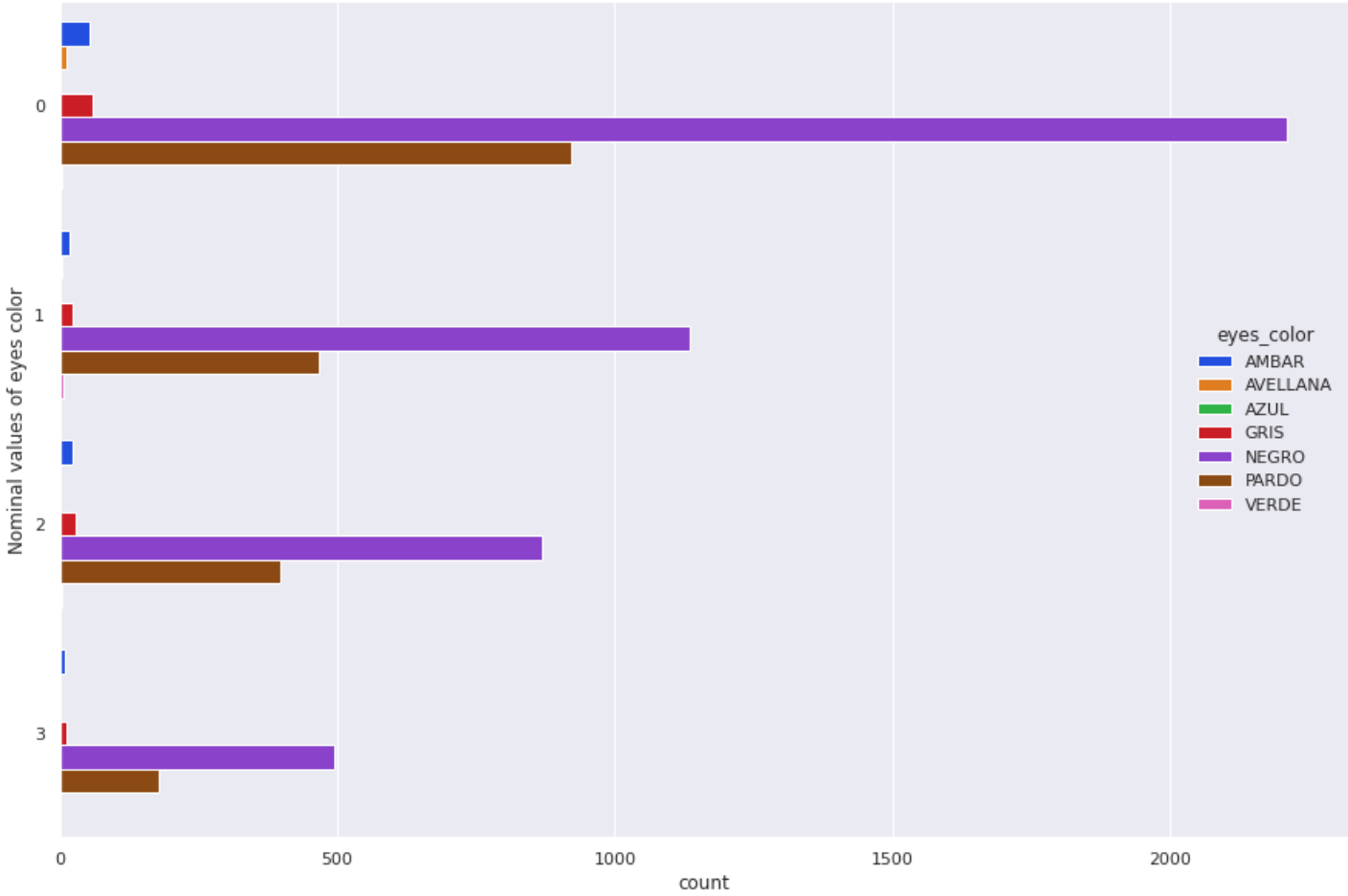


Figura 20

Diagrama de barras (Color de ojos x Edad).



IV.4.1.3.2. COLOR DE CABELLO

Dentro de los clústeres se observa que, para el color de cabello de los perfiles registrados, el valor "NEGRO" tiene la mayor cantidad de registros con casi el 86% y el valor "Azul" tiene la menor cantidad con 0.04%. También se observa que el Clúster 2 tiene el promedio de edad mayor y el Clúster 4 el promedio de menor.

Tabla 29

Resumen de distribución (Color de cabello x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
AZUL	1	0.01%	16	1	0.01%	17	1	0.01%	16	-	0%	-
GRIS	3	0.04%	14	4	0.06%	15.25	3	0.04%	16.33	-	0%	-
MARRON	438	6.32%	13.79	205	2.96%	14.11	165	2.38%	13.92	75	1.08%	13.79
NEGRO	2777	30.10%	13.81	1423	20.55%	13.84	1135	16.39%	13.82	618	8.92%	13.87
ROJO	10	0.14%	14.3	3	0.04%	16	8	0.12%	15.125	-	-	-
RUBIO	30	0.43%	14.27	11	0.16%	13.18	9	0.13%	13.11	5	0.07%	14.4
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 21

Diagrama de cajas (Color de cabello x Edad).

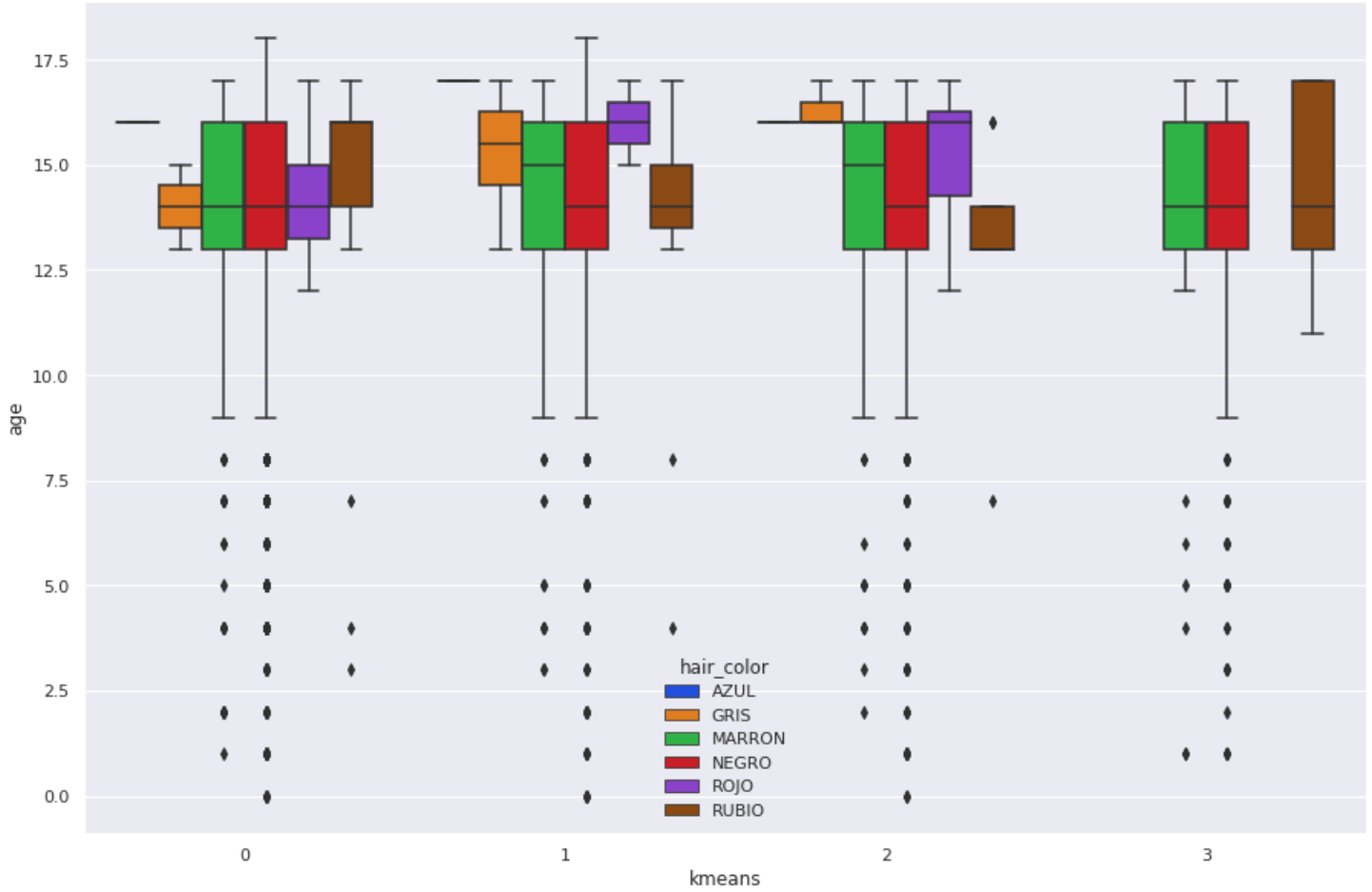
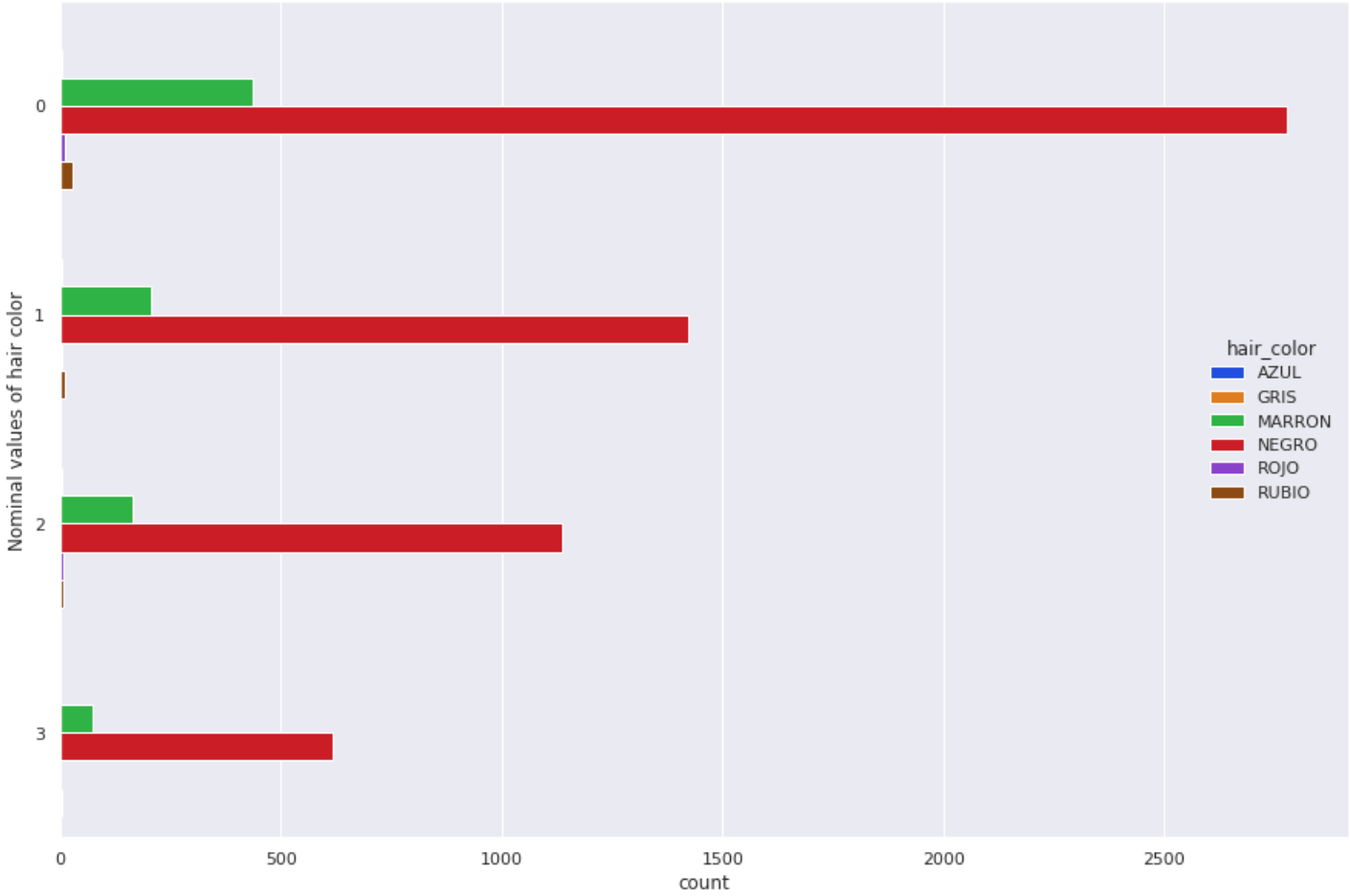


Figura 22

Diagrama de barras (Color de cabello x Edad).



IV.4.1.3.3. BOCA

Dentro de los clústeres para el atributo boca se observa que la mayor parte de registros tiene el valor de “MEDIANA” con el 68% del total. Además, podemos observar que el Clúster 3 tiene el promedio de edad menor y el Clúster 2 el mayor.

Tabla 30

Resumen de distribución (Boca x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
GRANDE	343	4.95%	14.45	185	2.47%	14.43	117	2.47%	14	74	1.07%	14
MEDIANA	2241	32.36%	13.99	1087	15.70%	14.04	924	13.34%	14.05	457	6.60%	13.88
PEQUEÑA	675	9.75%	12.92	375	5.42%	13.13	280	4.04%	13.1	167	2.41%	13.59
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 23

Diagrama de cajas (Boca x Edad).

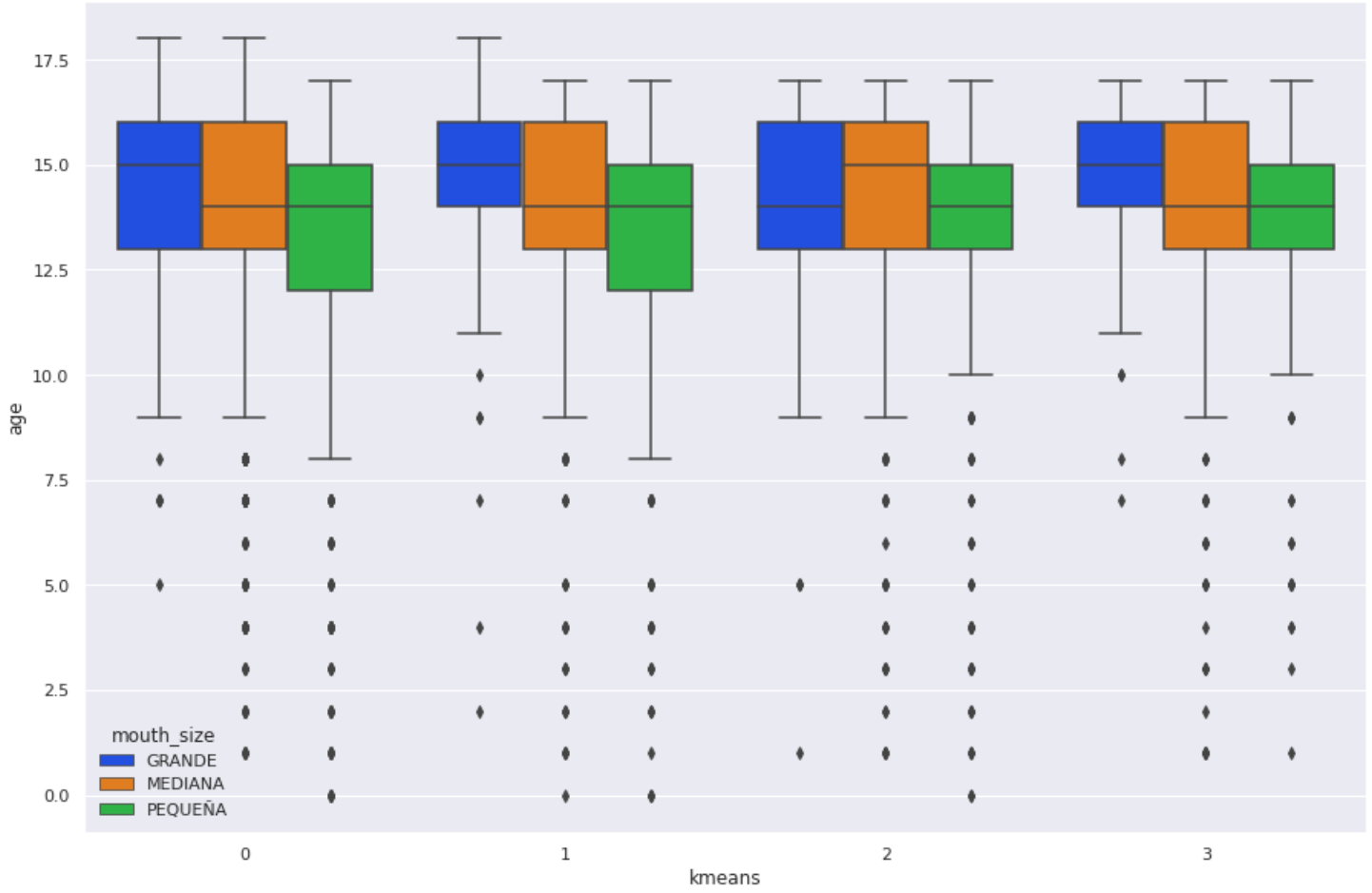
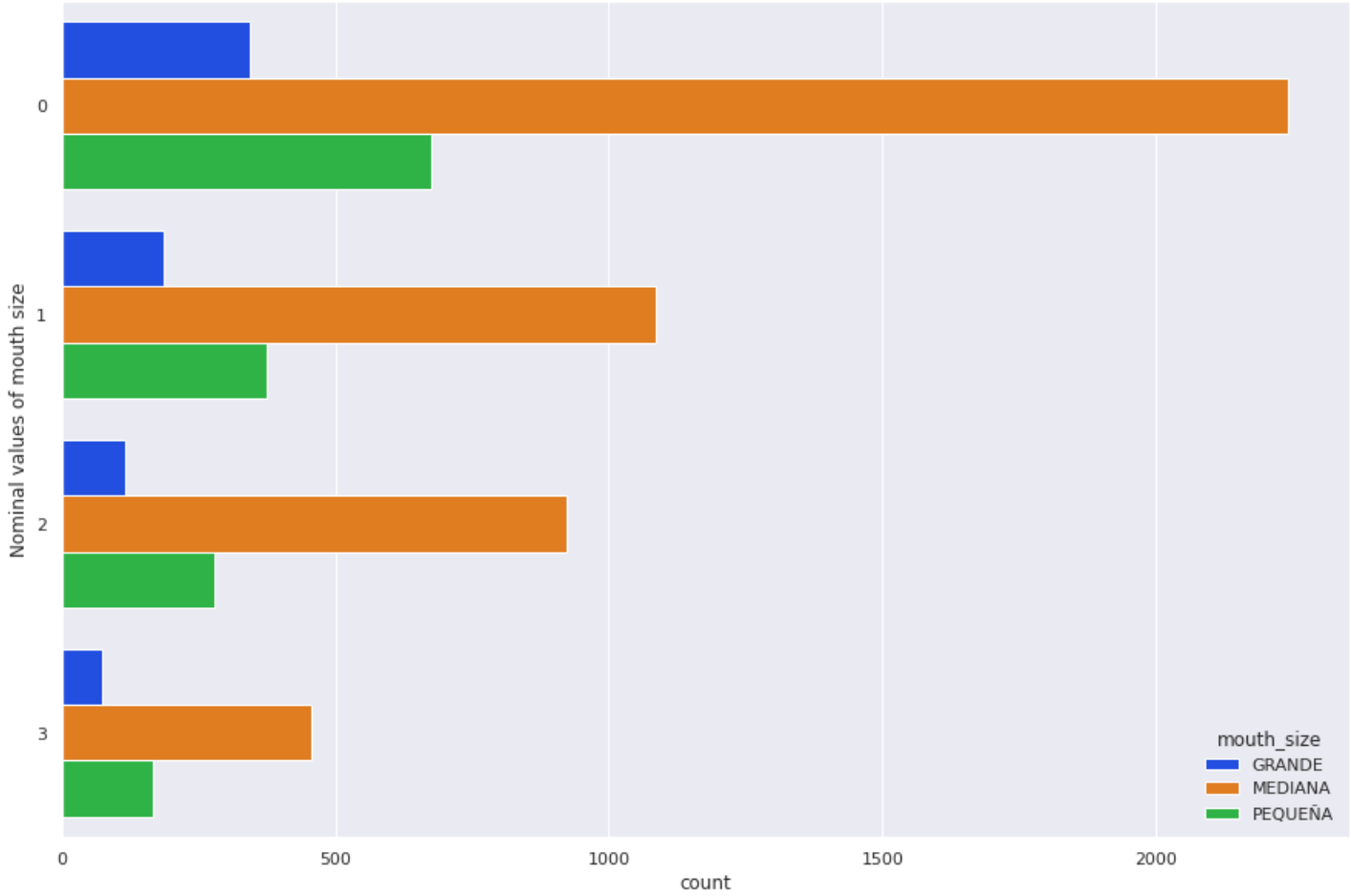


Figura 24

Diagrama de barras (Boca x Edad).



IV.4.1.3.4. NARIZ

Dentro de los clústeres podemos observar que la mayor cantidad de registros tiene el valor “MEDIANA” con casi el 76% del total. Además, El Clúster 4 tiene el promedio de edad menor y el Clúster 2 el mayor.

Tabla 31

Resumen de distribución (Nariz x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
GRANDE	441	6.37%	13.93	216	3.12%	14.24	151	2.18%	14.4	94	1.36%	13.95
MEDIANA	2458	35.49%	13.79	1237	17.86%	13.82	1040	15.02%	13.81	521	7.52%	13.89
PEQUEÑA	360	5.20%	13.86	194	2.80%	13.84	130	1.88%	13.52	83	1.20%	13.61
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 25

Diagrama de cajas (Nariz x Edad).

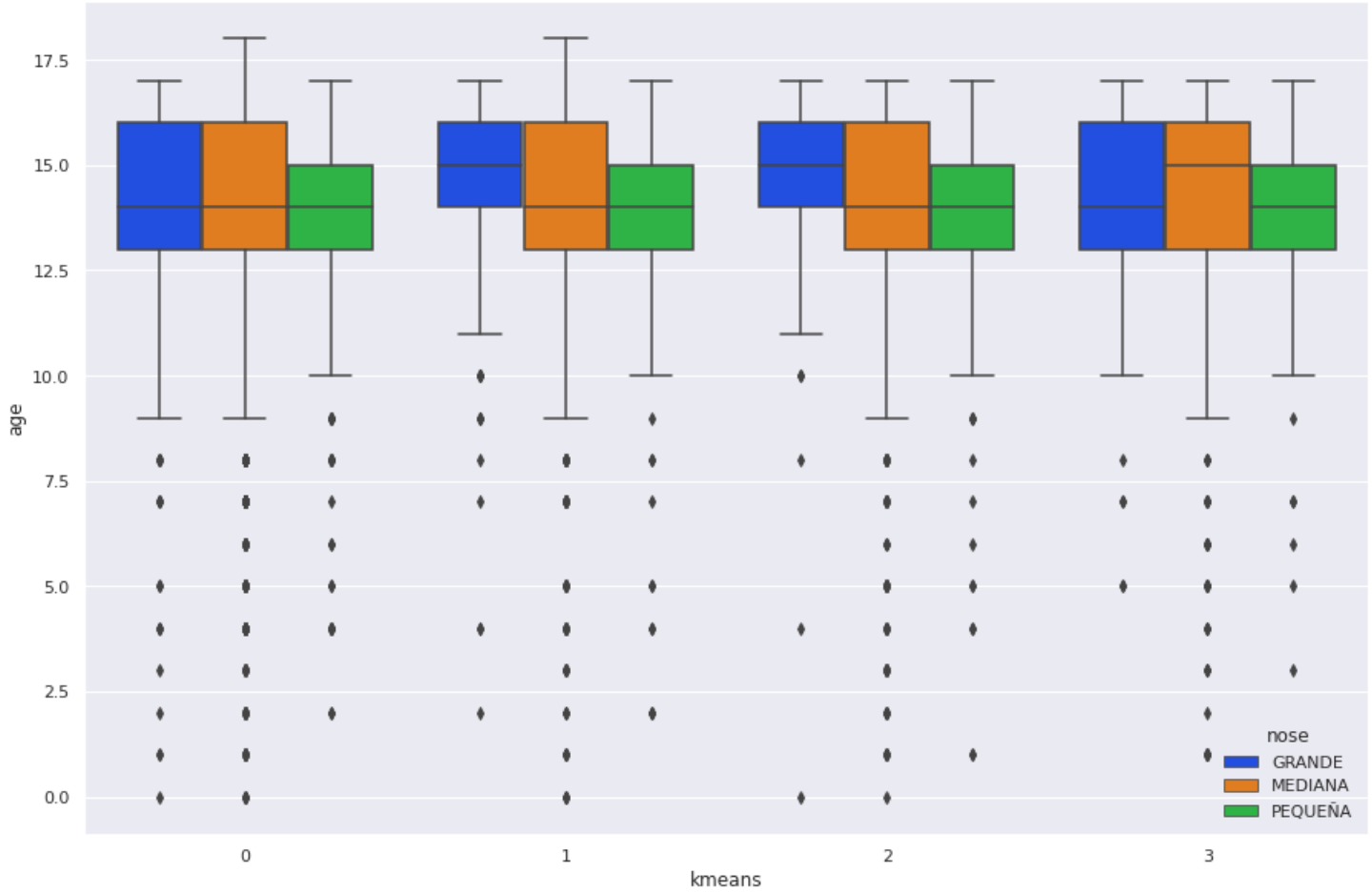
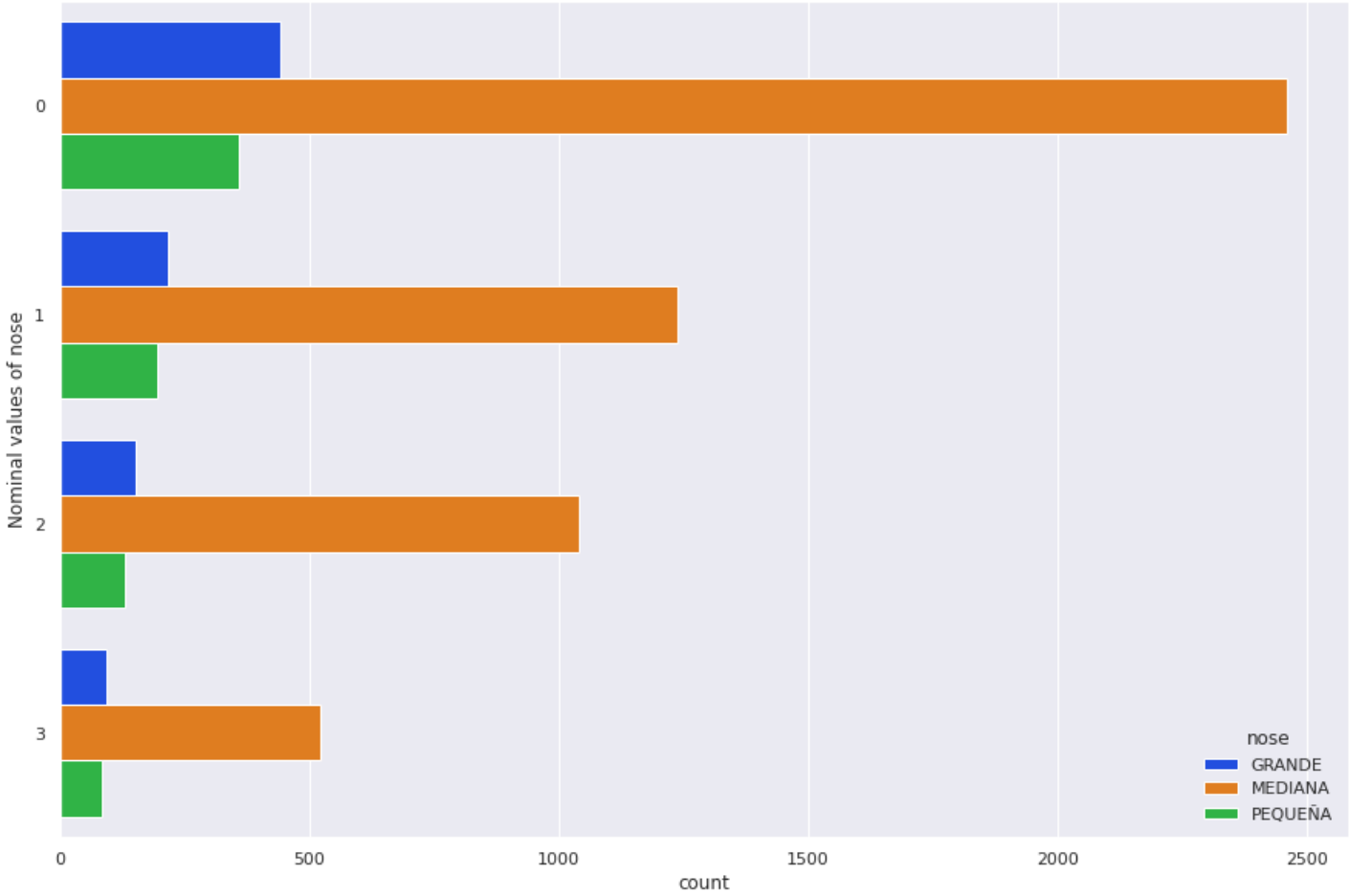


Figura 26

Diagrama de barras (Nariz x Edad).



IV.4.1.3.5. RAZA

Dentro de los clústeres podemos observar que para el atributo “raza” el valor con mayor cantidad es “MESTIZA” con poco más del 78%. Además, el Clúster 1 tiene el promedio de edad menor y el Clúster 2 el mayor.

Tabla 32

Resumen de distribución (Raza x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
BLANCA	365	5.14%	13.68	193	2.79%	13.84	176	2.54%	13.85	61	0.88%	13.66
MESTIZA	2581	37.17%	13.86	1271	18.35%	13.9	1004	14.50%	13.88	557	8.04%	13.94
NEGRA	12	0.17%	14.42	4	0.06%	15.25	8	0.12%	14.63	4	0.06%	14.5
TRIGUEÑA	310	4.48%	13.57	179	2.58%	13.72	133	1.92%	13.57	76	1.10%	13.46
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 27

Diagrama de cajas (Raza x Edad).

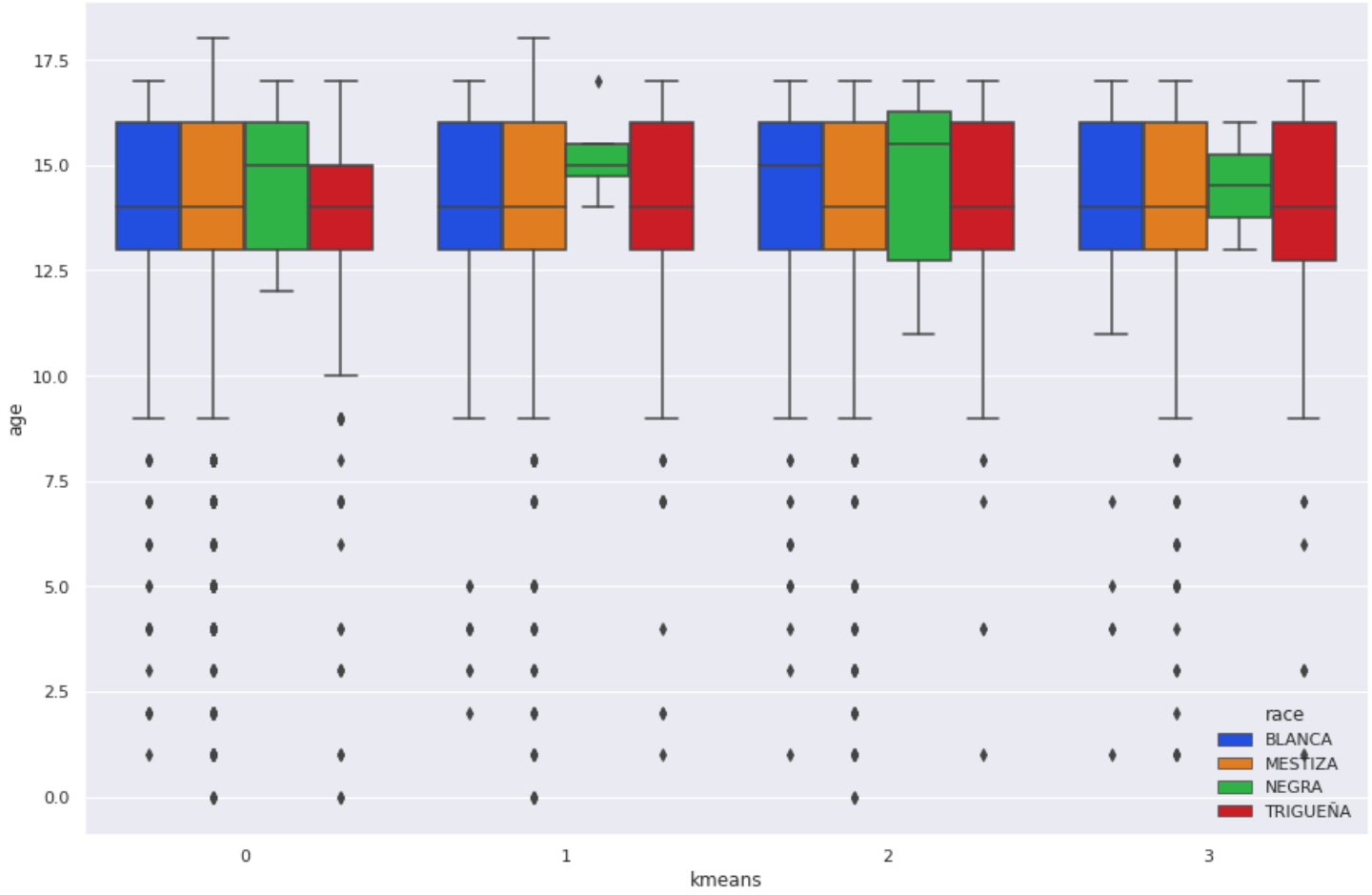
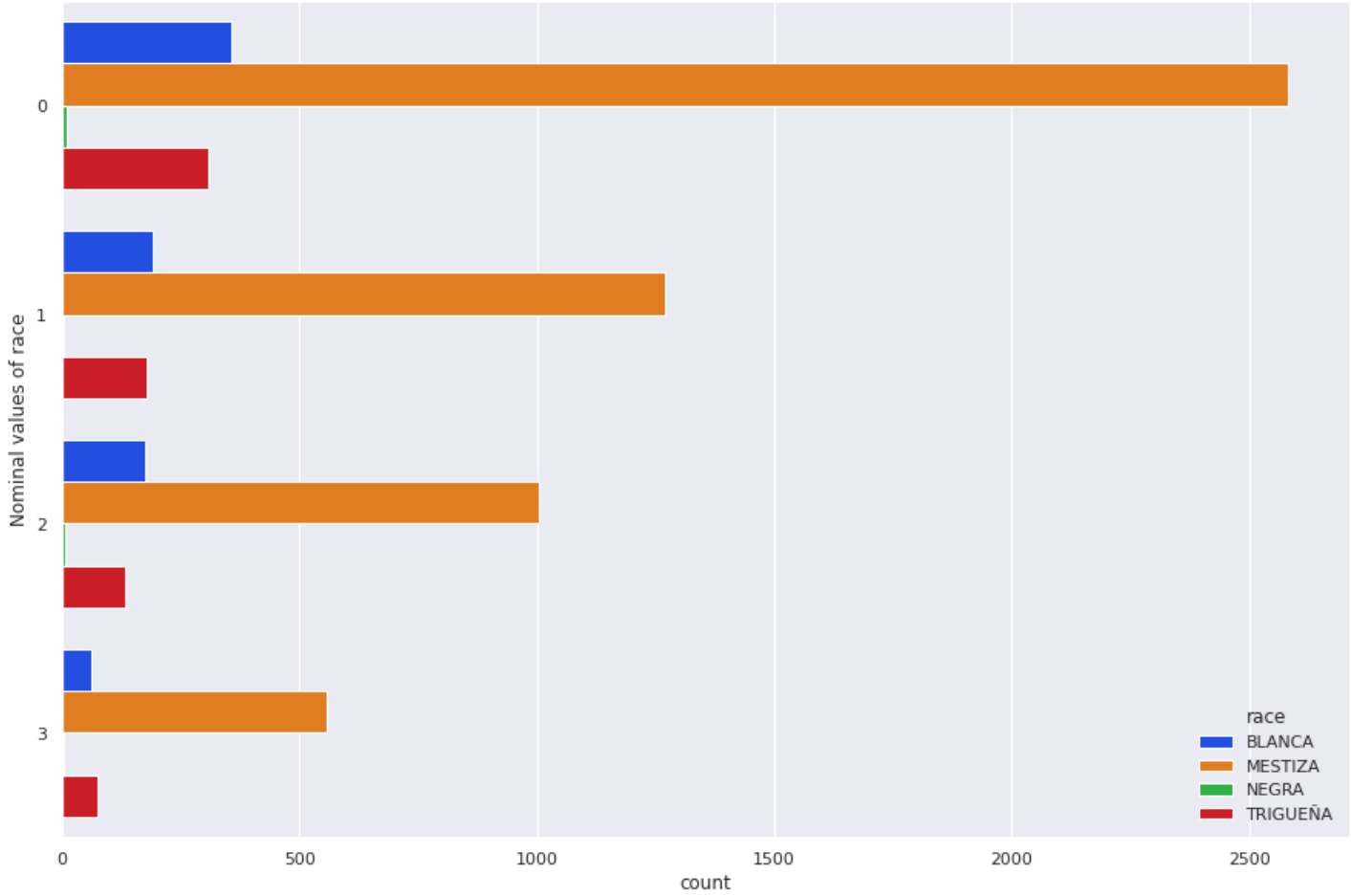


Figura 28

Diagrama de barras (Raza x Edad).



IV.4.1.3.6. GÉNERO

Podemos observar que la cantidad de niñas desaparecidas es casi 3 veces la cantidad de niños desaparecidos. También se ve que el Clúster 1 tiene el promedio de edad más bajo y el Clúster 4 el más alto.

Tabla 33

Resumen de distribución (Género x Edad).

Color de ojos	Clúster 1			Clúster 2			Clúster 3			Clúster 4		
	n	%	μ	n	%	μ	n	%	μ	n	%	μ
FEMENINO	2357	34.04	14.23	1177	17%	14.25	971	14.02%	14.2	526	7.60%	14.09
MASCULINO	902	13.03%	12.73	470	6.79%	12.96	350	5.05%	12.87	172	2.48%	13.17
Total	3259	47.06%		1647	23.78%		1321	19.08%		698	10.08%	

Figura 29

Diagrama de cajas (Género x Edad).

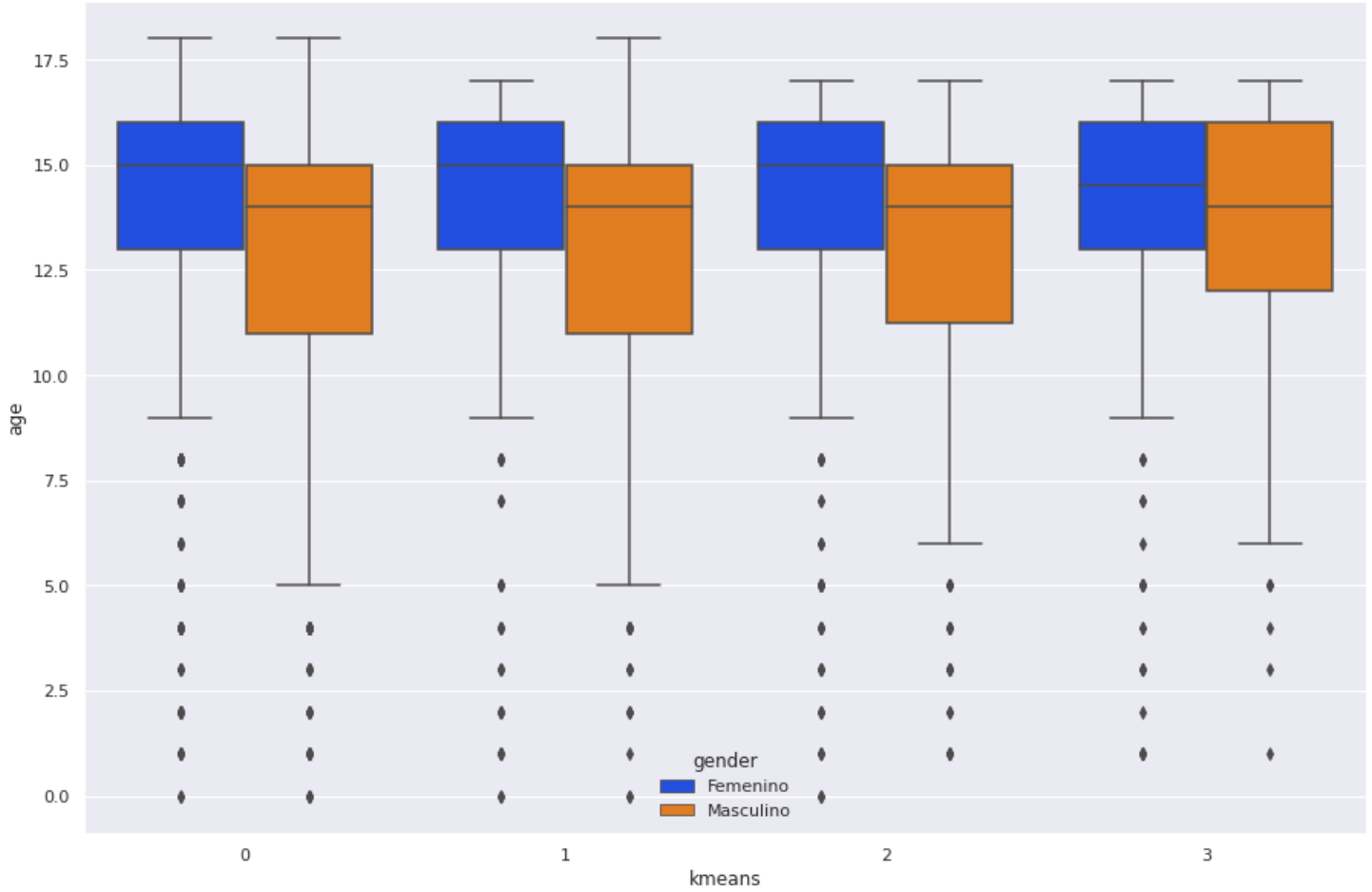
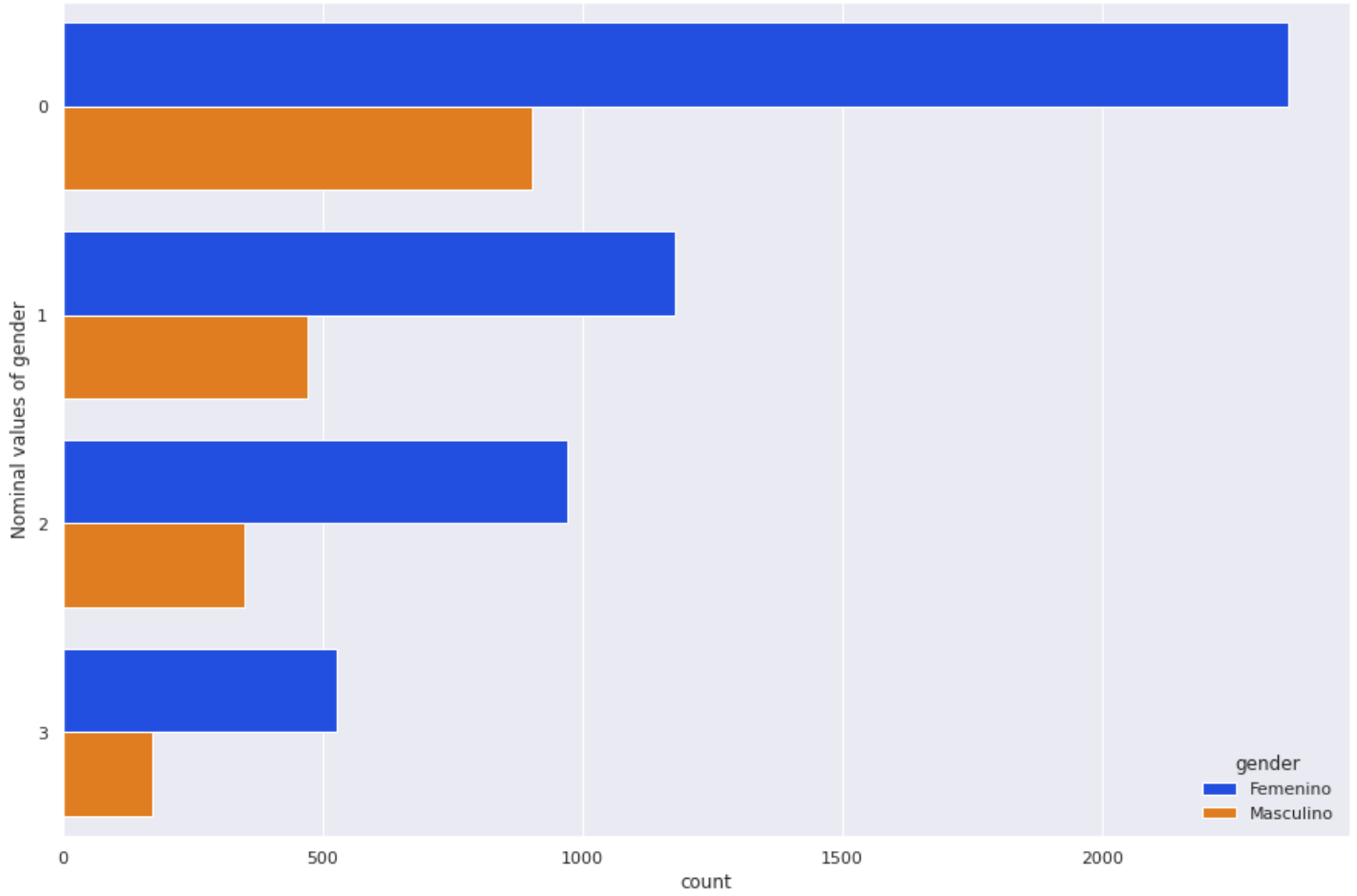


Figura 30

Diagrama de barras (Género x Edad).



CAPÍTULO V. DISCUSIÓN

En la presente investigación se realizó un estudio descriptivo que aplica técnicas de aprendizaje no supervisado para detectar patrones dentro del conjunto de datos de perfiles de menores desaparecidos a nivel nacional. La hipótesis planteada se valida con el criterio relativo detallado en el CAPÍTULO II, que propone comparar los resultados aplicados con un mismo algoritmo (k-means) variando los parámetros de entrada (número de clústeres).

K-means hace uso de la distancia euclidiana como medida de similitud entre registros, debido a esto los atributos deben de ser codificados a una representación numérica; la codificación resuelve las necesidades del algoritmo y permite que el proceso de clustering se termine en un tiempo corto, pero no representa adecuadamente al conjunto de datos.

El antecedente "ANALYZING AND CLUSTERING NEURAL DATA" se explora un conjunto de datos neurales obtenidos mediante Electrocorticografía, para esto fue necesario pasar los datos mediante filtros (Butterworth, Chebyshev) que servían como métodos de preprocesamiento de datos. Después se aplica un rango de valores aceptados para seleccionar los picos de variación asociados a la actividad cerebral. En la presente investigación se utilizaron otros métodos como la codificación binaria y el proceso KDD para ajustar los valores nominales a numéricos; si bien cada investigación requiere de técnicas de preprocesamiento exclusivos para el conjunto de datos debido a sus características, se observa que la aplicación de estas técnicas forma parte del análisis de clustering con el fin de obtener conocimiento.

Además, se aplicaron tres diferentes algoritmos de clustering para contrastar los resultados y determinar el algoritmo con mejores respuestas y que mantuviera los clústeres bien diferidos. Aunque en esta investigación no se realiza una comparación entre algoritmos, se optó por el algoritmo k-means que es el más usado en la literatura, debido a su velocidad para formar los clústeres y su eficacia para impedir la sobre posición de registros.

El antecedente "CLUSTERING ANALYSIS OF RESIDENTIAL LOADS", se realiza un análisis de clustering de 101 casas en Austin, Texas para observar el comportamiento de los consumidores según temporadas (verano, invierno, otoño y primavera) y contrastarlo con los datos de precios del mercado. Desde el comienzo de la investigación se tenía previsto la búsqueda de 3 clústeres debido a que en estudios previos se había encontrado 2, pero el investigador decidió agregar un clúster para disminuir la varianza de los subconjuntos.

Por lo tanto, no se utilizó ninguna técnica para determinar un numero adecuado de clústeres para proceder con la aplicación del algoritmo; al contrario del caso de esta investigación que no tiene un precedente en el campo de personas desaparecidas y se ve la necesidad de plantear un modelo de validación que permita determinar el valor adecuado mediante el uso de los índices y el método del codo.

El antecedente "CLÚSTER ANALYSIS OF CHILD HOMICIDE IN SOUTH KOREA" se analizan 341 casos de homicidio de niños en las edades entre 0 – 18 años, los datos registrados por caso tenían una variedad amplia de tipos; por lo cual se propuso utilizar la distancia Gower para calcular la similitud además del algoritmo PAM (Partición alrededor de medoides) para segmentar los clústeres adecuadamente, pero se menciona que esta estrategia solo es recomendada para conjuntos de baja numerosidad. La combinación de estas técnicas no se usa en esta investigación debido a la cantidad de registros (7612 perfiles) del conjunto de datos, debido al costo que representan estas técnicas en conjunto se optó por preprocesar los datos y usar el algoritmo k-means.

El antecedente "IMPLEMENTACIÓN DE UNA HERRAMIENTA DE ANÁLISIS DE RIESGO DE CRÉDITO BASADO EN EL MODELO DE RATING DE CRÉDITO, ALGORITMOS GENÉTICOS Y CLUSTERING JERÁRQUICO AGLOMERATIVO" propone una solución que combina dos estrategias de inteligencia artificial y demuestra un poder de predicción mayor comparado al modelo de regresión logística, además se menciona que debido a la metodología un experto humano puede interpretar fácilmente los resultados. En esta investigación se propone el uso del análisis de clustering para segmentar patrones del conjunto de datos, debido a que el resultado requiere ser interpretado por un humano experto.

El antecedente "APLICACIÓN DE LA MINERÍA DE DATOS DISTRIBUIDA USANDO ALGORITMO DE CLUSTERING K-MEANS PARA MEJORAR LA CALIDAD DE SERVICIOS DE LAS ORGANIZACIONES MODERNAS" se propone un algoritmo de clustering distribuido adaptable a la entidad judicial y concluye que la estrategia apoya en el cumplimiento de los objetivos de la entidad, lo cual mejora la calidad de sus servicios. Los pasos en el flujo de la propuesta no tienen la necesidad de adaptarse a las entidades involucradas en la gestión de los procesos que combaten la desaparición de menores, puesto que el estudio solo se enfoca en determinar y describir los patrones de los perfiles del conjunto de datos.

Sin embargo, los resultados de los patrones encontrados tienen la capacidad de mejorar los procesos para disminuir la desaparición de menores con ayuda de expertos.

GLOSARIO

1. **API:** Interfaz para programación de aplicaciones utilizado para acceder a funcionalidades de plataformas especificas.
2. **Binning:** Estrategia de suavizado de datos que propone separar los registros según a una agrupación definida.
3. **Clustering:** Técnica de aprendizaje no supervisado que busca agrupar registros de datos según a la distancia entre uno y otro.
4. **Comma-Separated Values (CSV):** Formato de archivo que separa valores usando comas.
5. **Dendrograma:** Representación grafica que mejora la visualización de subgrupos utilizando un esquema de árbol.
6. **Distancia Euclidiana:** Distancia entre dos puntos en el espacio euclídeo o espacio bidimensional.
7. **Método del codo:** Método grafico utilizado para visualizar números de clústeres asociados al costo de formación de clústeres.
8. **Error cuadrático medio:** resultado del calculo del promedio entre valores esperados y valores obtenidos, en un proceso de estimación.
9. **HyperText Markup Language (HTML):** Lenguaje de marcado de hipertexto utilizado comunmente para la elaboración de páginas web.
10. **HyperText Transfer Protocol (HTTP):** Protocolo de transferencia de hipertexto, utilizado en la capa de aplicación para transferir documentos de hipermedia.
11. **Knowledge Discovery from Data (KDD):** Proceso que propone tecnicas que ayudan a la formacion de conocimiento usando conjuntos de datos.
12. **K means:** Algoritmo de clustering que utiliza las medias como puntos de referencia para segmentar un conjunto de datos, hace uso de la distancia euclidiana y el proceso es constituido por múltiples iteraciones.

13. **Machine Learning:** Campo de la inteligencia artificial que busca entrenar algoritmos mediante la formación de conocimiento obtenido de los datos.

14. **Python:** Lenguaje de programación multiparadigma que se enfoca en la legibilidad del código.

15. **Web scraping:** Técnica que simula la interacción de un usuario con la web para recolectar datos.

CONCLUSIONES

- El proceso aplicado a los datos que aplico los pasos del proceso KDD y el análisis de clustering con el uso del algoritmo k-means mostro como resultado adecuado 4 patrones como se puede observar en el párrafo IV.3

- La herramienta de software diseñada y desarrollada para recolectar de datos de la página "Te Estamos Buscando" se basó en el concepto de web scraping, la herramienta permitió obtener los datos de perfiles de más de 7000 menores desaparecidos a nivel nacional.

- Dentro del conjunto de datos se registran múltiples atributos mencionados en la *Tabla 7*, de los cuales se seleccionaron 8 (altura, boca, color de cabello, color de ojos, edad, genero, nariz y raza) debido a que representan más a los individuos y conllevara a realizar una agrupación de acuerdo con sus características físicas.

- En el preprocesamiento de datos nos encontramos con diferentes pasos, estos son: integración, limpieza, transformación y reducción de datos. Debido a que el conjunto de datos posee atributos mixtos (binarios, nominales y numéricos) y el algoritmo de clustering seleccionado (k-means) calcula la similitud entre registros con la distancia euclidiana se requieren de valores numéricos para todos los atributos. Por lo tanto, en el proceso de transformación de datos se en codificaron los atributos no numéricos a representaciones binarias sintéticas lo que permitió adecuar los valores de las características y mejorar la eficiencia del algoritmo.

- Determinar el número adecuado de clústeres para procesar los datos puede ser una tarea muy complicada, debido a que pueden existir muchos factores que alteren los resultados. Por lo tanto, para validar la cantidad de clústeres que demuestra una segmentación eficiente del conjunto de datos, se utilizaron dos índices (Caliński y Harabasz, y Davies-Bouldin) los cuales demostraron tener mejores resultados cuando el número de clústeres es 4 cuyos resultados están en el punto 3 del capítulo IV.

RECOMENDACIONES

- Los registros de menores desaparecidos almacenan más propiedades a parte de las relacionadas al perfil de la persona, como las circunstancias de la desaparición o la vestimenta con la que fue visto o vista la última vez, por lo cual se recomienda una investigación más profunda tomando en consideración estos.

- La estrategia para detectar los patrones consistió en codificar los datos categóricos para utilizarlos con el algoritmo k-means lo cual nos dio los resultados detallados en el Capítulo V. Por otro lado, existen múltiples algoritmos de clustering y técnicas de procesamiento de datos, por lo cual se recomienda investigar estrategias con algoritmos y flujos de procesamiento más a fin a los tipos de datos mixtos.

- Esta investigación muestra los diferentes patrones que existen dentro del conjunto de datos con respecto a los perfiles de los menores, estos resultados pueden ser interpretados por expertos en campos de sociología, antropología, entre otros para determinar las razones de la desaparición de menores en el Perú.

REFERENCIAS

Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering. Algorithms and applications.* Chapman&Hall/CRC Data mining and Knowledge Discovery series.

Alpaydin, E. (2009). Introduction. En E. Alpaydin, *Introduction to machine learning* (págs. 1-20). MIT Press.

Amini, M.-R., & Usunier, N. (2015). *Learning with Partially Labeled and Interdependent Data.* Springer.

Amnistía Internacional . (14 de Septiembre de 2021). *Publicaciones - Las mujeres que nos faltan.* Obtenido de Amnistía Internacional : https://amnistia.org.pe/publicaciones/las-mujeres-que-nos-faltan/

Berndtsson, M., Hansson, J., Olsson, B., & Lundell, B. (2007). *Thesis Projects: a guide for students in computer science and information systems.* Springer Science & Business Media.

CIENCIACTIVA. (13 de Mayo de 2016). *CIENCIACTIVA.* Obtenido de CIENCIACTIVA: http://www.cienciactiva.gob.pe/images/bases/basica-y-aplicada/E041-2016-02-Bases-Integradas-del-Concurso.pdf

Davies, D. L., & W., B. D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 224-227.

El Peruano, Diario Oficial. (14 de Mayo de 2011). Normas legales. *El Peruano*, págs. 442436-442438.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques.* Morgan Kaufmann.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la Investigación.* Mexico: McGraw-Hill.

Internacional, A. (14 de Septiembre de 2021). *Publicaciones - Las mujeres que nos faltan.* Obtenido de Amnistía Internacional: https://amnistia.org.pe/publicaciones/las-mujeres-que-nos-faltan/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Application in R.* New York: Springer.

Jung, K., Kim, H., Lee, E., Choi, I., Lim, H., Lee, B., . . . Hong, H.-G. (2020). Cluster analysis of child homicide in South Korea. *Child Abuse & Neglect*, 104322.

Karimi, K. (22 de Abril de 2016). Clustering Analysis of Residential Loads. Manhattan, Kansas, Estados Unidos.

Kotu, V., & Deshpande, B. (2018). Clustering. En V. Kotu, & B. Deshpande, *Data Science: Concepts and Practice* (págs. 221-261). Morgan Kaufmann.

Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors.* Berlin: Springer.

Mamani Rodríguez, Z. E. (2015). Aplicación de la mineria de datos distribuida usando algoritmo de clustering k-means para mejorar la calidad de servicios de las organizaciones modernas. Lima, Lima, Perú.

Ministerio del Interior. (10 de 08 de 2021). *Desaparecidos en Perú*. Obtenido de Desaparecidos en Perú: https://desaparecidosenperu.policia.gob.pe/

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python.* Boston: O'Reilly Media.

Ramos Martinez, H. M. (2017). Implementación de una herramienta de análisis de riesgo de crédito basado en el modelo de rating de crédito, algoritmos genéticos y clustering jerárquico aglomerativo. Lima, Lima, Perú.

RENIPED, M. d. (9 de Octubre de 2021). *Registro Nacional de Información de Personas Desaparecidas.* Obtenido de Desaparecidos en Perú: https://desaparecidosenperu.policia.gob.pe/Desaparecidos/reniped

Sinha, A. (23 de Diciembre de 2015). Analyzing and clustering neural data. Boston, Massachusetts, Estados Unidos.

Skiena, S. S. (2017). *The Data Science Design Manual.* Springer.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition - Fourth Edition.*

Wittek, P. (2014). Unsupervised learning. En P. Wittek, *Quantum machine learning: what quantum computing means to data mining* (págs. 57-62). Academic Press.

Witten, I. H., Frank, E., & Hall, M. A. (2005). *Data Mining Practical Machine Learning Tools and Techniques.* Morgan Kaufman.

Xu, R., & Wunsch, D. C. (2008). *Clustering.* John Wiley & Sons.

ANEXOS

ANEXO A: DESCRIPCIÓN DE DATOS - CLÚSTER 1

Tabla 34

Descripción de atributos numéricos - Clúster 1.

	Edad	Altura (cm)
Recuento	3259	3259
Media	13.82	151.01
Desviación estándar	2.87	16.49
Mínimo	0	45
Máximo	18	193

Tabla 35

Descripción de atributos nominales - Clúster 1.

	Género	Color de ojos	Color de cabello	Boca	Nariz	Raza
Recuento	3259	3259	3259	3259	3259	3259
Valores únicos	2	7	6	3	3	4
Mayoría	Femenino	NEGRO	NEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia	2357	2210	2777	2241	2458	2581

Tabla 36

Distribución de valores de género - Clúster 1.

Valor	Cantidad
Femenino	2357
Masculino	902

Tabla 37

Distribución de valores de color de ojos - Clúster 1.

Valor	Cantidad
NEGRO	2210
PARDO	921
GRIS	59
AMBAR	53
AVELLANA	11
VERDE	4
AZUL	1

Tabla 38

Distribución de valores de color de cabello - Clúster 1.

Valor	Cantidad
NEGRO	2777
MARRON	438
RUBIO	30
ROJO	10
GRIS	3
AZUL	1

Tabla 39

Distribución de valores de boca - Clúster 1.

Valor	Cantidad
MEDIANA	2241
PEQUEÑA	675
GRANDE	343

Tabla 40

Distribución de valores de nariz - Clúster 1.

Valor	Cantidad
MEDIANA	2458
GRANDE	441
PEQUEÑA	360

Tabla 41

Distribución de valores de raza - Clúster 1.

Valor	Cantidad
MESTIZA	2581
BLANCA	356
TRIGUEÑA	310
NEGRA	12

ANEXO B: DESCRIPCIÓN DE DATOS – CLÚSTER 2

Tabla 42

Descripción de atributos numéricos - Clúster 2.

	Edad	Altura
Recuento	1647	1647
Media	13.88	151.50
Desviación estándar	2.74	14.99
Mínimo	0	50
Máximo	18	185

Tabla 43

Descripción de atributos nominales - Clúster 2.

	Género	Color de ojos	Color de cabello	Boca	Nariz	Raza
Recuento	1647	1647	1647	1647	1647	1647
Valores únicos	2	6	6	3	3	4
Mayoría	Femenino	NEGRO	MEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia	1177	1134	1423	1087	1237	1271

Tabla 44

Distribución de valores de género - Clúster 2.

Valor	Cantidad
Femenino	1177
Masculino	470

Tabla 45

Distribución de valores de color de ojos - Clúster 2.

Valor	Cantidad
NEGRO	1134
PARDO	466
GRIS	22
AMBAR	17
VERDE	5
AVELLANA	3

Tabla 46

Distribución de valores de color de cabello - Clúster 2.

Valor	Cantidad
NEGRO	1423
MARRON	205
RUBIO	11
GRIS	4
ROJO	3
AZUL	1

Tabla 47

Distribución de valores de boca - Clúster 2.

Valor	Cantidad
MEDIANA	1087
PEQUEÑA	375
GRANDE	185

Tabla 48

Distribución de valores de nariz - Clúster 2.

Valor	Cantidad
MEDIANA	1237
GRANDE	216
PEQUEÑA	194

Tabla 49

Distribución de valores de raza - Clúster 2.

Valor	Cantidad
MESTIZA	1271
BLANCA	193
TRIGUEÑA	179
NEGRA	4

ANEXO C: DESCRIPCIÓN DE DATOS – CLÚSTER 3

Tabla 50

Descripción de atributos numéricos - Clúster 3.

	Edad	Altura
Recuento	1321	1321
Media	13.85	150.95
Desviación estándar	2.89	15.87
Mínimo	0	50
Máximo	17	185

Tabla 51

Descripción de atributos nominales - Clúster 3.

	Género	Color de ojos	Color de cabello	Boca	Nariz	Raza
Recuento	1321	1321	1321	1321	1321	1321
Valores únicos	2	7	6	3	3	4
Mayoría	Femenino	NEGRO	MEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia	971	868	1135	924	1040	1004

Tabla 52

Distribución de valores de género - Clúster 3.

Valor	Cantidad
Femenino	971
Masculino	350

Tabla 53

Distribución de valores de color de ojos - Clúster 3.

Valor	Cantidad
NEGRO	868
PARDO	396
GRIS	29
AMBAR	22
VERDE	3
AVELLANA	2
AZUL	1

Tabla 54

Distribución de valores de color de cabello - Clúster 3.

Valor	Cantidad
NEGRO	1135
MARRON	165
RUBIO	9
ROJO	8
GRIS	3
AZUL	1

Tabla 55

Distribución de valores de boca - Clúster 3.

Valor	Cantidad
MEDIANA	924
PEQUEÑA	280
GRANDE	117

Tabla 56

Distribución de valores de nariz - Clúster 3.

Valor	Cantidad
MEDIANA	1040
GRANDE	151
PEQUEÑA	130

Tabla 57

Distribución de valores de raza - Clúster 3.

Valor	Cantidad
MESTIZA	1004
BLANCA	176
TRIGUEÑA	133
NEGRA	8

ANEXO D: DESCRIPCIÓN DE DATOS – CLÚSTER 4

Tabla 58

Descripción de atributos numéricos - Clúster 4.

	Edad	Altura
Recuento	698	698
Media	13.87	150.54
Desviación estándar	2.84	14.76
Mínimo	1	60
Máximo	17	184

Tabla 59

Descripción de atributos nominales - Clúster 4.

	Género	Color de ojos	Color de cabello	Boca	Nariz	Raza
Recuento	698	698	698	698	698	698
Valores únicos	2	6	3	3	3	4
Mayoría	Femenino	NEGRO	NEGRO	MEDIANA	MEDIANA	MESTIZA
Frecuencia	526	495	618	457	521	557

Tabla 60

Distribución de valores de género - Clúster 4.

Valor	Cantidad
Femenino	526
Masculino	172

Tabla 61

Distribución de valores de color de ojos - Clúster 4.

Valor	Cantidad
NEGRO	495
PARDO	178
GRIS	13
AMBAR	8
AVELLANA	2
VERDE	2

Tabla 62

Distribución de valores de color de cabello - Clúster 4.

Valor	Cantidad
NEGRO	618
MARRON	75
RUBIO	5

Tabla 63

Distribución de valores de boca - Clúster 4.

Valor	Cantidad
MEDIANA	457
PEQUEÑA	167
GRANDE	74

Tabla 64

Distribución de valores de nariz - Clúster 4.

Valor	Cantidad
MEDIANA	521
GRANDE	94
PEQUEÑA	83

Tabla 65

Distribución de valores de raza - Clúster 4.

Valor	Cantidad
MESTIZA	557
TRIGUEÑA	76
BLANCA	61
NEGRA	4