UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Una Institución Adventista

Sistema de Geo - sectorización de la inseguridad ciudadana para la sectorización de zonas delictivas en el contexto turístico utilizando Algoritmos de clustering.

Por:
Yuri Lisbeth Mamani Ramos

Asesor:
Ángel Rosendo Condori Coaquira

Juliaca, Julio del 2019

DECLARACIÓN JURADA
DE AUTORÍA DEL INFORME DE TESIS

Ing. Ángel Rosendo Condori Coaquira, de la Facultad de Ingeniería y Arquitectura, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que el presente informe de investigación titulado: **"SISTEMA DE GEO - SECTORIZACIÓN DE LA INSEGURIDAD CIUDADANA PARA LA SECTORIZACIÓN DE ZONAS DELICTIVAS EN EL CONTEXTO TURÍSTICO UTILIZANDO ALGORITMOS DE CLUSTERING"** constituye la memoria que presenta la bachiller Yuri Lisbeth Mamani Ramos para aspirar al título Profesional de Ingeniero de Sistemas ha sido realizada en la Universidad Peruana Unión bajo mi dirección.

Las opiniones y declaraciones en este informe son de entera responsabilidad del autor, sin comprometer a la institución.

Y estando de acuerdo, firmo la presente declaración en juliaca a los dos días del mes de julio del año dos mil diecinueve.

Ing. Ángel Rosendo Condori Coaquira.

Sistema de Geo - sectorización de la inseguridad ciudadana para la sectorización de zonas delictivas en el contexto turístico utilizando Algoritmos de clustering.

TESIS

Presentada para optar el título profesional de Ingeniero de Sistemas.

JURADO CALIFICADOR

Mg. Lennin Henry Centurión Julca
Presidente

Mg. Raúl Dante Gómez Apaza
Secretario

Ing. Ángel Rosendo Condori Coaquira
Asesor

Ing. Edér Gutierrez Quispe
Vocal

Juliaca, 02 de julio de 2019.

DEDICATORIA

Las personas lo suficientemente locas

como para pensar que pueden cambiar el mundo

son las que lo cambian.

Anuncio <<piensa diferente>> de apple, 1997.

AGRADECIMIENTO

Son muchas las personas que conforman mi formación académica, este trabajo es una forma de agradecer a mis padres que fueron modelos de lo que quería y no quería para mi futuro, y a mi madre, que me enseñó a luchar ante la adversidad y a creer en el (Dios).

Este trabajo fue financiado por el CONCYTEC FONDECYT en el marco de la convocatoria "Proyecto Investigación Básica 2015" [número de contratación 184-2015].

Sin embargo, las personas más directamente responsables de esta tesis incluyen a la doctora Gladys Maquera que siempre fue y es una fuente de inspiración a lo no convencional y común la cual confío en mi persona para poner un granito de arena en su proyecto del CONCYTEC y la Universidad Peruana Unión "PLATAFORMA DIGITAL INTELIGENTE Y BIG DATA PARA EL TURISMO RURAL COMUNITARIO EN LA REGIÓN PUNO" del cual aprendí mucho y me llevó no solo un aprendizaje para mi carrera y vida, sino una manera muy distinta de ver la realidad de un profesional, y también por el financiamiento dado mediante el proyecto por parte de las dos instituciones, también darle gracias a mi asesor ingeniero Ángel Rosendo Condori Coaquira que fue una guía importante para la construcción y ejecución de la tesis, a mis compañeros que fueron fuente de inspiración.

Pero principalmente agradezco a Dios por permitirme formarme, conocer y disfrutar a personas tan grandiosas y espectaculares, por haberme impulsado a ser fuerte y nunca haberme dejado a la deriva.

ÍNDICE GENERAL

DEDICATORIA iv
AGRADECIMIENTO v
ÍNDICE DE TABLAS ix
ÍNDICE DE FIGURAS x
ÍNDICE DE ANEXOS xii
ABREVIATURAS Y ACRÓNIMOS xiii
RESUMEN xiv
ABSTRACT xv
CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA 15
1.1. Definición del problema 15
1.2. Justificación 16
1.2.1. Justificación social 16
1.2.2. Justificación económica 17
1.2.3. Utilidad Teórica 18
1.2.4. Utilidad práctica 20
1.3. OBJETIVOS 21
1.3.1. Objetivo General 21
1.3.2. Objetivos específicos 21
CAPÍTULO II: BASES TEÓRICAS 22
2.1. Revisión de la literatura 22
2.2. Seguridad 24
2.2.1. Seguridad ciudadana 24
2.2.2. Seguridad turística 26
2.3. Inteligencia artificial 28
2.3.1. Lenguaje natural de procesamiento 30
2.3.2. Técnicas de agrupamiento y reconocimiento de patrones. 30
2.3.3. Técnicas de agrupamiento para datos cualitativos y cuantitativos 31
2.4. Minería de datos de datos 33
2.4.1. Base de datos para la minería de datos 33
2.5. Minería de datos 34
2.6. Web scraping y minería de datos 37
2.6.1. Maquetación web 37

2.6.2. Extracción de datos 38
2.6.3. Tipos de web scraping 38
2.6.4. Tipos de datos 39
2.7. Estudio de datos 39
2.7.1. Python para ciencia de datos 39
CAPÍTULO III: MATERIALES Y MÉTODOS 41
3.1. Descripción de lugar de ejecución 41
3.2. Metodología de la investigación 41
3.2.1. Tipo de investigación 41
3.2.2. Investigación propositiva 41
3.2.3. Investigación aplicada 42
3.2.4. Arquitectura de solución 42
3.3. Herramientas tecnológicas 42
3.4. METODOLOGÍA 44
3.4.1. Metodología para la gestión de proyectos 44
3.4.2. Metodología de aplicación CRISP-DM 47
3.4.3. Desarrollo de los objetivos 49
3.4.4. Objetivo 01: Recolectar datos estructurados y no estructurados con técnicas de web scraping. 49
3.4.5. Objetivo N° 2: Analizar los datos recolectados con Python para ciencia de datos y Weka. 53
Limpieza de datos estructurados 55
3.4.6. Objetivo N° 3: Preparar los datos y aplicar algoritmos de clustering a los datos recolectados. 56
3.4.7. Objetivo N° 4: Diseñar la plataforma para la visualización de datos. 58
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN 60
4.1. Resultados de los objetivos. 60
4.1.1. Objetivo 01: Recolectar datos estructurados y no estructurados con técnicas de web scraping. 60
4.1.2. Objetivo N° 2: Analizar los datos recolectados con Python para ciencia de datos y Weka. 61
4.1.3. Objetivo N° 3: Preparar los datos y aplicar algoritmos de clustering a los datos recolectados. 66
4.1.4. Objetivo N° 4: Diseñar la plataforma para la visualización de datos. 72
CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES 73
5.1. Conclusiones 73

5.2. Recomendaciones 73
REFERENCIAS BIBLIOGRÁFICAS 75
ANEXOS 77
Anexo 1: Estadística de la WEF 77
Anexo 2: Mapic del proyecto de investigación 78

ÍNDICE DE TABLAS

Tabla 1 Etapas del proceso de Data Mining.....36
Tabla 2 Estructura de web scraping.....38
Tabla 3 Diferenciación de los tipos de datos 39
Tabla 4 Librerías de Python 40
Tabla 5 Herramientas de extracción de datos. 42
Tabla 6 Herramientas d estudio de datos.....43
Tabla 7 herramientas de visualización de datos 43
Tabla 8 herramientas de implementación de datos 44
Tabla 9 Variables para estudio 51
Tabla 10 Variables de extracción de datos. 52
Tabla 11 Variables de estudio. 54
Tabla 12 Variables e identificación de atributos de los datos.....55
Tabla 13 Delitos y faltas.....64

ÍNDICE DE FIGURAS

Figura 1 Porción del turismo PBI total....18
Figura 2 Estadística de las principales causas de inseguridad en el Perú....25
Figura 3 Almacenamiento estratégico del plan nacional de seguridad ciudadana....26
Figura 4 Estadística del PBI del banco mundial del Perú....27
Figura 5 Áreas de aplicación de la Inteligencia Artificial. 29
Figura 6 Mapa mental de ramas o sub áreas de la Inteligencia Artificial....29
Figura 7 Arquitectura de un sistema experto.....30
Figura 8 Datos agrupados 31
Figura 9 Formula agrupación de medición cuántica por Span y Content.....31
Figura 10 Especificación de variables Span y Content. 32
Figura 11 Formula cualitativa basada en distancias de centroides de Gowda y Diday. 32
Figura 12 Especificación de las variables de la fórmula de Gowda y Diday.....32
Figura 13 Algoritmo de clustering para datos cualitativos. 32
Figura 14 Lista de algoritmos de Clasificación de agrupamiento algoritmos. 33
Figura 15 Categorías de algoritmos de minería de datos.....35
Figura 16 Procesos de KDD de minería de datos.....37
Figura 17 Clasificación de las técnicas de Data Mining 37
Figura 18 Arquitectura de solución al problema.....42
Figura 19 Principios de OpenUp....45
Figura 20 Elementos del OpenUp 45
Figura 21 Ciclo de vida de OpenUP 46
Figura 22 Fases de OpenUP 46
Figura 23 Modelo de proceso CRISP–DM ([CRISP-DM, 2000]). 48
Figura 24 Variables de datos Data Base....50
Figura 25 Base de datos....51
Figura 26 Página de extracción de datos 53
Figura 27 Código de extracción de datos.....53
Figura 28 . Weka y atributos de los datos....54
Figura 29 Weka y atributos de los datos....55
Figura 30 data con atributos originales 56
Figura 31 Data limpia....56
Figura 32 Entorno de desarrollo....56
Figura 33 Datos a estudio con JUPYTER....57
Figura 34 resultados de comuna.....57
Figura 35 Estructura de módulos.....58
Figura 36 Prototipo de visualización.....59
Figura 37 Datos extraídos de la página nube.....60
Figura 38 Resultados de weka análisis de clustering.....61
Figura 39 Incidencias delictivas.....61
Figura 40 Centroides de incidencias. 62
Figura 41 Tipos delitos clasificados.....62
Figura 42 Análisis de los cluster de tipo delito.....62
Figura 43 Resultados de comuna en Python.....63
Figura 44 Comuna resultados de combinaciones.....63
Figura 45 Comuna gráfica de barras....63

Figura 46 hurto sin violencia. gráfica de barras. 64
Figura 47 hurto sin violencia gráfica 65
Figura 48 hurto sin violencia R combinaciones. 65
Figura 49 Barrios gráfica de barras. 66
Figura 50 Barrios combinaciones gráfica de barras. 66
Figura 51 Geo Sectorización de datos 66
Figura 52 Barra y geo sectorización. 67
Figura 53 Geo sectorización por puntos en delito. 67
Figura 54 Comuna gráfica de barras y geo sectorización 67
Figura 55 barrios según delitos geo sectorización. 68
Figura 56 Días y horas. 68
Figura 57 gráfica de barras de tipos de delitos. 69
Figura 58 gráfica de barras T_D 69
Figura 59 Estructura de datos. 70
Figura 60 K medias gráfica para centroides 70
Figura 61 K centroides. 71
Figura 62 gráficas de clusters según 4 centroides 71
Figura 63 Prototipado del sistema de geo sectorización 72

ÍNDICE DE ANEXOS

Anexo 1 Estadística de la WEF.....77
Anexo 2 Mapic del proyecto de investigación.....78

ABREVIATURAS Y ACRÓNIMOS

- PENTUR: Plan estratégico Nacional de turismo.
- IA: inteligencia Artificial.
- PNP: Policía Nacional del Perú.
- K means: Algoritmo de clasificación.
- PBI: Producto interno Bruto.
- KDD: KDD: Knowledge Discovery in Databases.
- WEF: World Economic Forum.
- OMT: Organismo Mundial del Turismo.
- MINCETUR: Ministerio de comercio exterior.
- REACH: Aplicación de denuncias.
- GP: algoritmo genético.
- TIC: Tecnologías de información y telecomunicaciones.
- PMI: Project management institute.
- XP: Programación extrema.
- DIRTEPOL: Dirección Territorial policial.
- TIP: tipos de datos.
- INEI: Instituto nacional de estadística e informática
- CONASEC: Seguridad ciudadana.
- BBVA: Banco Bilbao Viscaya Argentaria.
- WTTC: World Travel & Tourism Council
- SE: Sistemas expertos.
- SBC: Sistemas basados en conocimientos.
- API: Application Programming Interface.

RESUMEN

El propósito de la investigación es diseñar y analizar un sistema de geo sectorización con técnicas de Ciencia de Datos e IA, utilizando datos históricos de delitos. El agrupamiento con el algoritmo de K-Means de las grandes cantidades de datos los estudios de las variables a agrupar son propias del giro de negocio (turismo) del problema de la inseguridad ciudadana en el contexto turístico. Para la construcción del prototipo del sistema se adoptó dos metodologías: OpenUp para el desarrollo y CRISP-MD para el procesamiento de datos e integración del algoritmo. Para el desarrollo se recolectó datos estructurados y semi estructurados, los datos estructurados fueron tomados a estudio del repositorio público de la ciudad de buenos aires ya que tiene las variables similares a utilizar del PNP del dep. turismo, los datos semi estructurados fueron extraídos mediante técnicas de web scraping de lugares turísticos posteriormente analizados con la herramienta weka y Python utilizado las técnicas de ciencia de datos, Así mismo se trabajó con el algoritmo de clustering K Means en donde se obtuvo gráficos de agrupamiento de datos de acuerdo al centroide principal de las variables: delito y lugar, en Power BI obtuvimos la visualización basada en grafica de mapas. El resultado fue el análisis de las variables en construcción para el turismo, en el cual obtuvimos los datos estadísticos y el algoritmo de agrupamiento, construyendo según los análisis el prototipo del sistema para el sector turismo y la complejidad tanto para la integración con el algoritmo y su forma de alimentación al sistema, para así poder validar el sistema web/móvil desplegada en sus servidores de aplicación.

Palabras clave: Clustering, visualización de datos, algoritmos, ciencia de datos y seguridad turística.

ABSTRACT

The purpose of the research is to design and analyze a geo-sectorization system with data science and AI techniques, using historical crime data. The grouping with the algorithm of K-Means of the large amounts of data the study of the variables to be grouped are typical of the business (tourism) turn of the problem of citizen insecurity in the tourism context. For the construction of the system prototype, two methodologies were adopted: OpenUp for the development and CRISP-MD for the data processing and integration of the algorithm. For the development, structured and semi-structured data was collected, the structured data were taken to study the public repository of the city of Buenos Aires, since it has similar variables to use from PENTUR-Peru, the semi-structured data were extracted using web techniques scraping of tourist places later analyzed with the tool weka and python used the techniques of science of data, Likewise it was worked with the algorithm of clustering K Means where graphics of grouping of data according to the main centroid of the variables was obtained: crime and place, in Power BI we obtained the visualization based on map graphics. The result was the analysis of the variables under construction for tourism, in which we obtained the statistical data and the grouping algorithm, building according to the analysis the prototype of the system for the tourism sector and the complexity both for the integration with the algorithm and its way of feeding the system, in order to validate the web / mobile system deployed in its application servers.

Keywords: Clustering, data visualization, algorithms, data science and tourism security.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1. Definición del problema

La seguridad turística es una de las debilidades del país que se encuentra en el puesto 108 del ranking del Country/Economy Profiles World Economic Forum (WEF, 2017). "Actualmente nuestro país ha avanzado del puesto 58 en el 2015 al puesto 51 en el 2017 en la industria del turismo en materia de viajes y experiencia turística, descubriendo el camino para un crecimiento más sostenible e inclusivo, destaca la fidelidad y compromiso con la industria de viajes y turismo, mejorando su seguridad y protección creando entornos más propicios para las empresas turísticas o afines y desarrollar su infraestructura para mejorar la conectividad" (World Economic Forum WEF, 2017) (Anexo 1). "La percepción de seguridad ciudadana por parte de la mayoría de peruanos es calificada como baja o nula, esto debido a que la delincuencia se ha convertido en el principal tema de preocupación tanto de los ciudadanos mismos como de las autoridades" (Fernández y Rivas, 2015).

Para la Organización Mundial del Turismo (OMT, s.f) "El desarrollo económico concebido para mejorar la calidad de vida de la comunidad receptora o emprendedores que buscan el desarrollo turístico, mantener la calidad del medio ambiente y de la seguridad turística del que dependen tanto la comunidad anfitriona como los visitantes". Citado por Travel (2013). Se busca alcanzar la eficiencia en tecnología de producción en temas de Turismo es una de las necesidades primordiales en la competencia empresarial y uno de los intereses y problemas más frecuentes en el cumplimiento de tal meta, en cuanto al tiempo que puede abarcar los resultados y más aún si hablamos del desarrollo del país en temas de economía albergando a la población y sus recursos naturales.

El estudio y la visualización de datos en los últimos años ha tomado gran importancia, debido al crecimiento de los datos, los mismos que representan un nuevo recurso para diversas áreas de investigación, ya que permite obtener información que sea

interpretada mediante el desarrollo de un software. Así como también las compañías buscan atender necesidades de un público específico, Una buena forma de buscar la gran cantidad de información es a través de dimensiones geográficas asociadas a los datos, ya que la información georreferenciada se hace más necesaria ya que una de las formas de analizar la información es a través de filtros (Ubicación) permitiendo obtener información de interés, dependiendo de donde se encuentra el consumidor de la información.

Pimentel (2016), "Actualmente en el mercado existen varios recursos para la presentación de datos georreferenciados, entre ellos se encuentran las herramientas de Google Maps y MapBox. Ambas compañías han hecho un gran trabajo generando mapas y servicios para que otros desarrolladores puedan trabajar sobre ellos creando nuevas herramientas". De Acuerdo a los datos de inseguridad ciudadana y la utilización de diferentes herramientas, se requiere la georreferenciación y el agrupamiento de los datos, ya que se rescata con ello los propósitos de descubrir el estudio y análisis de los datos, buscando la clasificación natural de los datos mediante las similitudes entre ellos. Es por ende que en este estudio nace la problemática relacionada al estudio geográfico de zonas con escenarios de inseguridad turística: **¿Cómo facilitar el manejo de la información en respuesta a la resolución de eventos delictivos los cuales requieren velocidad de procesamiento de acuerdo a la cantidad de datos? Cómo brindar información de utilidad a los turistas: ¿Visualización de indicadores de criminalidad en diferentes zonas?**

1.2. Justificación

1.2.1. Justificación social

El estudio de datos y la implementación de un sistemas capaz de sectorizar las zonas de alto nivel de hechos de inseguridad turística tendrá un impacto no solo para la sociedad peruana sino para los extranjeros que visiten los lugares turísticos, recursos turísticos,

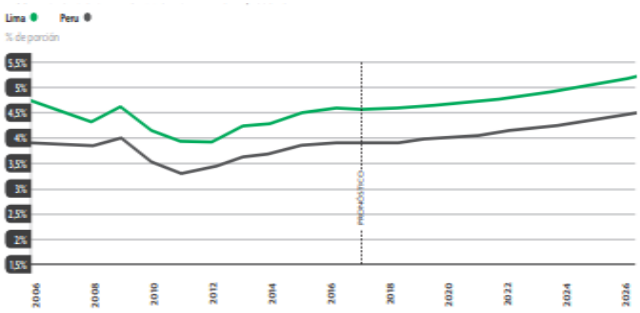
servicios turísticos y otros agentes involucrados en el turismo peruano, además de beneficiar directamente a las instituciones como lo son la policía nacional, hospitales, etc. otorgando un mejor servicio de las instituciones encargadas del tema y así causar un impacto favorable para el turismo peruano.

(Carlos J, 2018) "El turismo al ser una actividad sensible ante cualquier situación que se note inseguridad, el sector requiere del Estado en su conjunto la estabilidad política que permita reactivar la economía, la estabilidad jurídica que exigen los empresarios e inversionistas, y la estabilidad social que garantice a los turistas un viaje sin contratiempos por territorio patrio. Que los hechos ocurridos en el 2017 –desastres naturales, conflictos sociales, incertidumbre electoral, crisis política, etc.–, marquen la pauta de lo que debemos y no hacer este año, lo que hay que prever y cómo debemos actuar a futuro. El Perú merece recuperar su imagen de país en crecimiento y destino seguro para los turistas".

1.2.2. Justificación económica

Según el (banco mundial en Perú, 2018) "Nuestro país Perú es uno de los países de américa latina que tiene potencial turístico a desarrollar tanto como recurso natural y/o materia de producción. La diferenciación del crecimiento de millones de personas en las diferentes áreas de la economía dentro de un país, generando trabajos en la población, reduce la pobreza y desarrolla la industrialización, es netamente la responsabilidad del estado brindar oportunidades de crecimiento. El cual mide el conjunto de factores y políticas que permite el desarrollo sostenible del sector turismo en cuanto a la competitividad en el crecimiento del país en materia de riqueza económica albergando los recursos naturales. Dando crecimiento económico a las empresas de turismos, agencias públicas y privadas, Lugares y recursos naturales, servicios de turismos y a la comunidad en general".

"En el último año el Perú ha logrado desarrollar un crecimiento potencial en la economía con una cifra de 4,6% en contribución directa al PBI en el sector de viajes y turismo, siendo una de las cifras más altas en américa latina" según el reporte económico de World travel & tourism council del sector de viajes y turismo en las ciudades (2017) américa latina la cual se representa en la Figura 1.



	Resumen de Lima				Resumen de Perú			
			Crecimiento promedio				Crecimiento promedio	
	2016	2026	2006-16	2016-26	2016	2026	2006-16	2016-26
PIB VyT (US$ mm)	45	90	8,9%	7,2%	7,6	15,6	8,3%	7,5%
PIB total (%)	4,6%	5,2%	-	-	3,9%	4,5%	-	-
Empleo VyT (000)	181,6	248,4	2,2%	3,2%	403,5	566,5	2,1%	3,5%
Empleo total (%)	3,3%	3,7%	-	-	2,5%	2,9%	-	-

Figura 1 Porción del turismo PBI total

Fuente: World travel & tourism council.

El ministerio de comercio exterior y turismo (MINCETUR, 2017) "Ha propuesto el plan de protección al turista 2017 - 2018 ya que es una herramienta de gestión dedicada a mejorar la condición de seguridad con una perspectiva de largo alcance, la cual garantiza al turista una estadía libre de riesgo. Ya que con ello se beneficiaría las instituciones encargadas del orden de turismo y empresas, organizaciones etc. involucradas en el turismo peruano".

1.2.3. Utilidad Teórica

El estudio de los datos de criminalidad si está dada en el Perú como lo es el caso de una aplicación peruana premiada por el príncipe del reino unido el cual fue desarrollado por el joven Moisés Salazar Vila el cual es autodidacta, el cual fue reconocido como el profesional peruano más valioso de Microsoft, la aplicación es conocida como REACH

lanzada como tal en 2014 la aplicación está dedicada a reportar las actividades de diferentes incidencias, así como también está siendo optimizada para que pueda ser implementada con algún algoritmo de inteligencia artificial, biometría, reconocimiento artificial, Big Data para mejorar aún más la herramienta pues cuando llegue al millón de usuarios no podrá darle el soporte necesario él solo. Así como también las tesis de estudios que se realizó para el distrito de molina utilizando minería de datos el cual fue desarrollado con el título de "sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de la molina utilizando minería de datos" por Jorge Julio Jaulis Rúa en 2015, el cual tiene como objetivo mejorar el proceso de prevención de delitos mediante la minería de datos. Otro caso es el estudio de los datos de acuerdo a la criminalidad basado en la seguridad ciudadana que se realizó en la tesis de "sistema de mapeo digital de zonas delictivas utilizando un algoritmo genético progresivo" realizada por Br. Buddy Richard Oruna Rodríguez, que permite mejorar la elaboración de mapas digitales utilizando algoritmos GP que permite la identificación de zonas con mayor índice delictivo fue desarrollada en Trujillo, Perú 2015. Por otra parte, los estudios dirigidos al turismo en el contexto de seguridad están ampliamente ligada a la seguridad ciudadana el cual no se encontró estudios en el tema Turístico a nivel de Perú, así como también no se encontró en la región de puno. En la región de puno no se ha dado el estudio de datos como Minería de datos o inteligencia artificial que componen la carrera de ingeniería de sistemas a nivel de criminalidad o estudio de datos de sectorización de zonas delictivas en el contexto ciudadano o turístico. Por otra parte, cabe mencionar la aplicación que lanzó el MINCETUR (Police Tourist) que sólo proporciona la facilidad de alertar a la policía, serenazgo, hospitales y bomberos de las incidencias más cercanas para un pronto auxilio, pero no cuenta con algún estudio de datos o sectorización de zonas delincuenciales o de diferentes incidencias que puedan atravesar los turistas.

En los últimos años se habla de que no es posible hablar de desarrollo sin hablar de las Tecnologías de Información y Comunicación (TIC) más aún si las tecnologías son inteligentes y pueden ayudar al desarrollo de alguna organización o simplifique los esfuerzos, porque se considera que el acceso a la información y la creación de nuevos conocimientos forman una parte esencial en los procesos de desarrollo. Así como también la seguridad debe ser interpretada como un estado subjetivo que nos permite percibir que nos desplazamos de un espacio turístico exento de riesgos reales o potenciales. La percepción de inseguridad afecta negativamente a la experiencia turística e impacta negativamente en la imagen del destino dando una solución tecnología inteligente el cual pueda dar información de los lugares y sus estados de inseguridad. Mejorar los niveles de seguridad turística mediante alianzas estratégicas con los actores del sector público y privado en beneficio de la imagen del país y el desarrollo socioeconómico del Perú.

(Rodriguez, 2017). “Para la etapa de predicción se realiza un agrupamiento o “clustering” de la nueva información (nuevos eventos delictivos), incluidos en la etapa temporal, con la finalidad de encontrar los posibles centros de las zonas de riesgo (a tractores criminales). Una vez que se tienen los centros, por medio de un modelo empírico de un a tractor criminal y con una determinada ventana de tiempo, se genera una distribución de predicción del riesgo criminal. En la actualidad, con las nuevas herramientas de localización es posible obtener información geo-referenciada de los probables sectores, los mismos que se usan para el proceso de encontrar patrones espacio-temporales de los incidentes, así se sabrá cuándo y dónde un nuevo crimen puede ocurrir”.

1.2.4. Utilidad práctica

La policía nacional del Perú: En el cual ayudará a la sectorización zonas delictivas más propensas a surgir los delitos, de acuerdo al análisis que pueda ser mostrado en los datos

estudiados. Este caso de estudio servirá también para las instituciones como lo son Serenazgo, Municipalidades, entre otros agentes encargados de la seguridad turística, tanto nacional, regional y local.

MINCETUR: Los cuales podrán ahondar en el plan de protección al turista el cual busca proteger los intereses y satisfacer la estancia del turista, en el cual podrá respaldar con bases de análisis de datos la confiabilidad de zonas seguras de los recursos, así poder brindar mayor seguridad y una experiencia agradable al turista. así como también encontramos beneficiadas las instituciones, los emprendedores, agencias y todas aquellas empresas con giros de negocio en el turismo.

1.3. OBJETIVOS

1.3.1. Objetivo General

Analizar y diseñar un prototipo de Sistema de Geo - sectorización de la inseguridad ciudadana para la sectorización de zonas delictivas en el contexto turístico utilizando Algoritmos de clustering.

1.3.2. Objetivos específicos

- ➢ Recolectar datos estructurados y no estructurados con técnicas de web scraping.
- ➢ Analizar los datos recolectados con Python para ciencia de datos y Weka.
- ➢ Preparar los datos y aplicar algoritmos de clustering a los datos recolectados.
- ➢ Diseñar la plataforma para la visualización de datos.

CAPÍTULO II: BASES TEÓRICAS

2.1. Revisión de la literatura

Ignacio Perversi,2007 Aplicación de minería de datos para la exportación y detección de patrones delictivos en argentina. En el objetivo, de este trabajo es realizar una implementación de minería de datos en el análisis de información criminal de información criminal en Argentina y comprobar sus efectividad y valor agregado.

Jorge Julio Jaulis Rua, 2015. Sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de la molina utilizando minería de datos, Objetivo mejorar el rendimiento del proceso de prevención del delito de las comisarías de la molina utilizando Minería de datos. La metodología de clasificación supervisada basada en Redes Bayesianas, aplicando la metodología de minería de datos y la metodología de gestión de proyectos PMI - PBOOK. El modelo de minería de datos que recoge información histórica de todo el año 2015 de las denuncias registradas en cada una de las comisarías, y en base a un algoritmo de aprendizaje automático arroja las zonas más propensas a la ocurrencia de algún hecho delictivo, para ello se optó por mostrar esta información a través de mapas de la zona y que esto puedan ser accedidos desde cualquier dispositivo. Como resultado se obtuvo la mejora de procesos y la implementación de un sistema de predicción de hechos delictivos en el distrito de la molina.

Buddy Oruna Rodriguez 2015, Sistema de mapeo digital de zonas delictivas utilizando un algoritmo genético progresivo, Objetivos mejorar y reducir la elaboración de un sistema de mapeo digital de zonas delictivas a través de un algoritmo genético progresivo en la unidad de identificación de la oficina criminalística. Utilizando algoritmos genéticos con métodos adaptativos que utiliza la estructura de los algoritmos genéticos progresivos, Desarrollado con la metodología de XP, El cual resuelve los problemas de búsqueda y

optimización, teniendo como resultado un sistema de mapeo digital de zonas delictivas y la identificación de la oficina de criminalística de la III DIRTEPOL.

Juan Rodriguez 2017, Desarrollo de una metodología para caracterizar y predecir el riesgo criminal mediante la generación de espacio - temporales empíricos basados en manejo de datos. El objetivo general es generar una metodología basada en la relación entre eventos criminales y locación de servicios que proporcione un modelo espacio-temporal que permita caracterizar y predecir zonas de riesgo criminal dentro de un área determinada. La metodología empleada para la caracterización y predicción del riesgo criminal. La metodología propuesta para la etapa de predicción fue usar el algoritmo K-means para agrupar los nuevos eventos criminales introducidos en la actualización del modelo dinámico, de esta forma se encuentran los centros de las posibles zonas de riesgo criminal. Con la idea de "a tractor criminal" se generó un modelo empírico dinámico, con el cual los eventos criminales fueron atraídos según la cercanía al a tractor. Analizando estos resultados, se puede apreciar cómo se evalúa cada uno de los modelos, según el número de datos generados. El TIP resultante es alto para algunos resultados de la evaluación, llegando en algunos casos a valores de TIP muy cercanos a uno, siendo un buen resultado en la evaluación del modelo de predicción generado. Los resultados se obtuvieron evaluando el modelo generando varias distribuciones de predicción con diferentes ventanas de tiempo.

Tania Denisse Gonzalez Villa 2013. Análisis, diseño e implementación de un sistema web y móvil para el soporte informático a la gestión de los servicios de atención que brinda las comisarías de la comunidad. Que tiene como objetivos gestionar eficientemente los procesos que soportan algunos de los servicios que brinda una comisaría y proporcionar información para la seguridad de ciudadanos. El cual sigue la metodología de aplicada para la gestión de proyecto SCRUM y la metodología aplicada para el desarrollo del producto XP. Que tiene como resultado la tabla comparativa del estado de arte de la solución, solución

“SeguriApp” Implementada en la plataforma Android integrada con los servicios de Google Maps, la tecnología de Realidad Aumentada y los servicios de Facebook y Twitter y solución “SeguriApp” Implementada en la plataforma Web integrada con los servicios de Google Maps y la tecnología cometD

2.2. Seguridad

Para la organización de las naciones unidas (ONU, s. f) “La seguridad humana subraya las necesidades de poner énfasis a lo que vendría hacer los programas de la paz y la seguridad, el desarrollo y los derechos humanos de manera la cual sea más eficiente, eficaz y orienta a la prevención”.

2.2.1. Seguridad ciudadana

(Salas, 2016). “La seguridad es un asunto público, ya que forma parte de la convivencia cotidiana de los ciudadanos en un marco de integridad y salvaguarda los derechos de las personas es por ello la importancia del significado de espacio público. La seguridad también es un bien común o público; es indivisible que debe proveerse de manera imparcial y al ser una condición de interés social en todo el territorio nacional”.

“Durante el último semestre (noviembre 2017-abril 2018), el 25,5% de la población de 15 y más años de edad a nivel nacional (urbano) fue víctima de algún hecho delictivo, según el Instituto Nacional de Estadística e Informática” INEI (2018).

Para la (INEI, 2018). “Las principales causas de inseguridad ciudadana es la débil participación de los ciudadanos, sociedad civil, sector privado y medios de comunicación. en la cual se suma la falta de articulación de estrategias entre el estado y medios de comunicación, así como la pérdida de valores, cultura cívica y respeto a la ley. Escasos

espacios públicos seguros como lugares de encuentro ciudadano con una baja cultura de la población en la conservación de los espacios públicos y en las reglas de convivencia urbana".

Figura 2 Estadística de las principales causas de inseguridad en el Perú

Fuente: Dirección general de seguridad ciudadana.

Plan nacional de seguridad ciudadana (conasec, 2013) "Conjuntamente con la política pública con objetivos alineados a la política general del estado al 2021 (Bicentenario), que tiene como objetivos estratégicos: El sistemas nacional de seguridad ciudadana, la implementación de espacios seguros, la reducción de factores de riesgo social que favorecen en comportamientos delictivos, promover la participación de los ciudadanos, la sociedad civil, el sector privado y los medios de comunicación para enfrentar la inseguridad ciudadana, así como también fortalecer a la policía nacional del Perú como una institución moderna, con una gestión eficaz, eficiente y con altos niveles de confianza ciudadana, en tanto mejorar el sistema de la administración de la justicia para la reducción de la delincuencia".

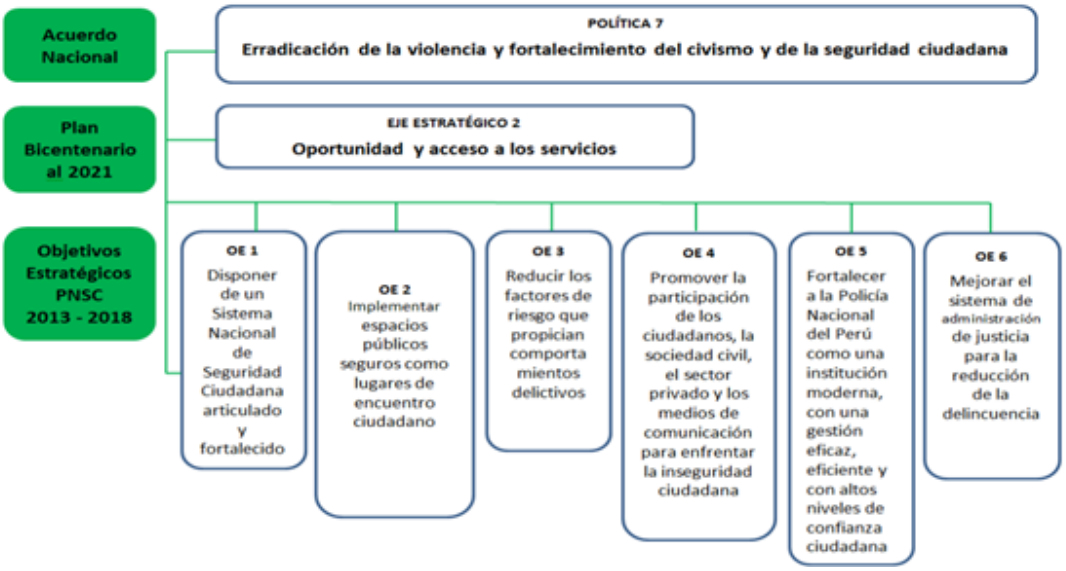


Figura 3 Almacenamiento estratégico del plan nacional de seguridad ciudadana

Fuente: Dirección general de seguridad ciudadana.

2.2.2. Seguridad turística

La seguridad turística es uno de los pilares para que extranjeros puedan tener la seguridad de tener una bonita experiencia, es por lo cual el Ministerio de comercio exterior y turismo ha elaborado la red de protección al turista 2017 - 2018, Plan de protección al turista. (MINCETUR, 2017). “El Plan de Protección al Turista, ha sido elaborado como resultado del trabajo consensuado entre los diversos actores públicos y privados, tomándose como base el trabajo desarrollado por el Ministerio de Comercio Exterior y Turismo – MINCETUR, el Ministerio del Interior – MININTER, la Dirección de Turismo PNP en el ámbito de la Seguridad Turística, en el marco de la Ley N° 28982 – Ley que regula la Protección y Defensa del Turista, la Ley N° 29408 - Ley General de Turismo y los lineamientos estratégicos del Plan Estratégico Nacional de Turismo – PENTUR. el cual tiene como objetivo principal mejorar los niveles de seguridad turística mediante alianzas estratégicas con los actores del sector público y privado en beneficio de la imagen país y el desarrollo socio-económico del Perú”.

MINCETUR (2017) “Manifiesta que generar las condiciones favorables para que los turistas nacionales y extranjeros, así como las comunidades receptoras, puedan desarrollar sus actividades turísticas en un entorno libre de riesgos físicos, materiales y sicológicos, es el propósito principal del presente Plan, constituyendo un reto su aplicación a partir de un trabajo interinstitucional e interinstitucional. Gestionar óptimamente la Seguridad Turística en el Perú es contribuir al fortalecimiento de nuestra competitividad turística”.

La seguridad conlleva a la permanencia del turista libre de riesgo y una experiencia inolvidable en el territorio peruano, así aportar una variable para la suma en la elección de un destino turístico y sostenibilidad turística. según BBVA Research Perú (2018) “los turistas extranjeros gastan siete veces más que los nacionales El estudio de BBVA Research identifica que el viajero del extranjero destina alrededor de USD 1,000 por viaje, siete veces más que un turista interno. El informe precisa que el gasto promedio por viaje de los turistas nacionales, tomando como base información de 2015, asciende a S/ 451, incluyendo el transporte. En tanto, los extranjeros destinan unos US$ 994 por viaje, sin considerar el costo del pasaje de ingreso ni salida del país”, viéndose la Figura 4.

Figura 4 Estadística del PBI del banco mundial del Perú

Fuente: Banco Mundial del Perú.

La contribución del turismo a la economía global mantendrá su dinamismo, El World Travel & Tourism Council (WTTC, 2015) "Señala que el crecimiento del sector turismo será más dinámico en comparación con otras actividades económicas. Se espera que, hacia el 2025, el viaje y turismo proporcionen 72,9 millones de nuevos puestos de trabajo, de los cuales 23,2 millones se generarán dentro del sector. Asimismo, la contribución del PBI total del viaje y el turismo a la economía en general aumentará del 8,8% en el 2014 al 10,5% en el 2025, mientras que el empleo del 9,4% al 10,7%. La clave de este aumento radicará en el crecimiento esperado de la demanda de los mercados emergentes, en los cuales se identifica una creciente proporción del gasto del consumidor al viaje y el turismo".

2.3. Inteligencia artificial

La inteligencia artificial es una potencia intelectual, que tiene la facultad de conocer o entender, la cual es aplicado a máquinas que tienen un conocimiento o comprensión de un proceso a realizar, las cuales pueden realizar una o varias de las capacidades estudiadas los cuales son sistemas expertos (SE) y sistemas basados en conocimiento (SBC). Según uno de los pioneros de la IA Marvin Minsky "la inteligencia artificial es una ciencia de construir máquinas para que hagan cosas que, si las hicieran los humanos, requerirían inteligencia".

El mayor problema que enfrenta esta disciplina es el análisis de cómo los humanos buscan soluciones y resuelven los innumerables problemas. Áreas de aplicación:

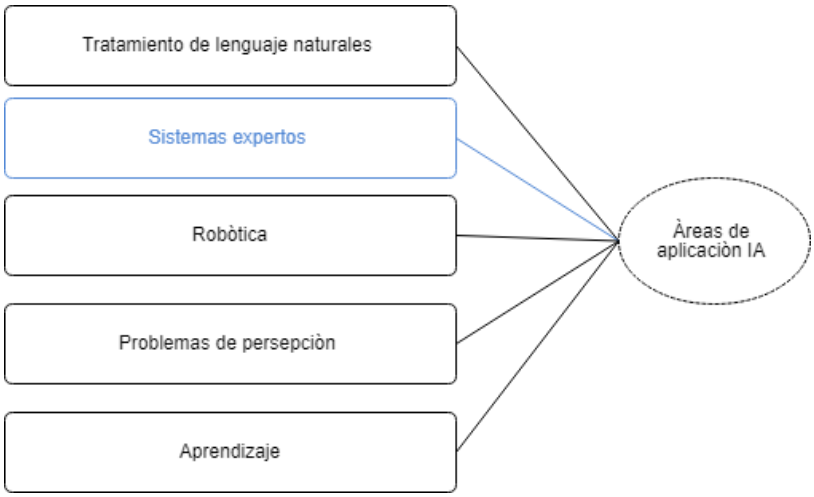


Figura 5 Áreas de aplicación de la Inteligencia Artificial.

Fuente: Elaboración propia, estudiando los esquemas de IA.

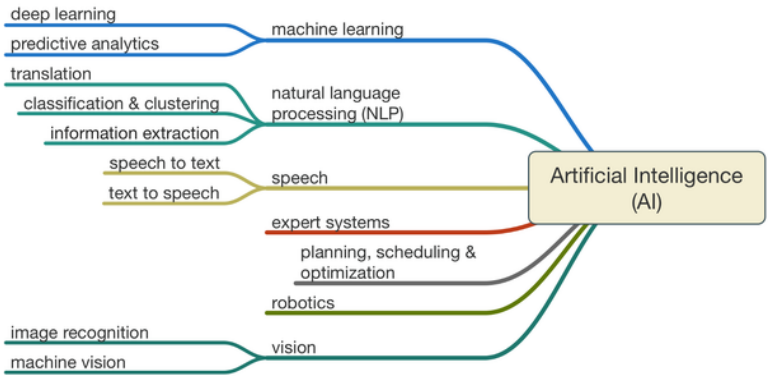


Figura 6 Mapa mental de ramas o sub áreas de la Inteligencia Artificial.

Fuente: AI Mind Map – Machine Learning and Artificial Intelligence Study Group – Medium.

Los sistemas expertos están englobados en todos aquellos sistemas en donde las informaciones pueden conseguirse reducciones más cercanas a la realidad. Los sistemas expertos se encargan de tareas como la resolución del problema de la misma forma que del ser humano, trabajar datos de información dudosa o incompletos, explicar resultados obtenidos, aprender conocimientos nuevos sobre los datos y restauración de conocimiento. Los resultados obtenidos de estos sistemas expertos son más fáciles de documentar.

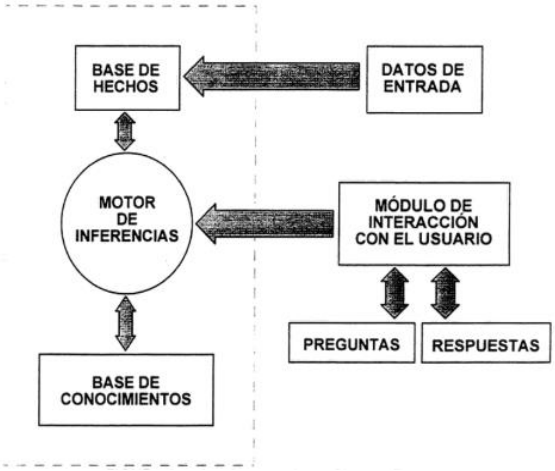


Figura 7 Arquitectura de un sistema experto.

Fuente: Editorial Servicios de Publicaciones Universidad de Oviedo Introducción a la ingeniería Artificial: Sistemas Expertos, Redes Neuronales Artificiales y Computación Evolutiva.

2.3.1. Lenguaje natural de procesamiento

(Escolano, F., Cazorla, M. A., Alfonso, M. I., Colomina, O., & Lozano, 2003) "Este tipo de lenguaje es el que nos permite el designar las cosas actuales y razonar acerca de ellas, fue desarrollado y organizado a partir de la experiencia humana y puede ser utilizado para analizar situaciones altamente complejas y razonar muy sutilmente".

2.3.2. Técnicas de agrupamiento y reconocimiento de patrones.

Las tecnologías de información han dado un gran impulso estudiando las teorías y técnicas de reconocimiento implantadas en distintos sistemas de información. En la cual se maneja información representada en forma de patrones complejos: textos escritos, numero, música flores, etc.

(Benitez y Diez, 2005). "El funcionamiento de los algoritmos de clustering están basados en la optimización de una función objetivo a resolver con su respectivo proceso a analizar y respuesta a dar, que normalmente es la suma ponderada de las distancias a los centros, aunque estas funciones pueden variar, y muchas veces los distintos algoritmos de reconocimiento de patrones se distinguen principalmente en la definición de sus funciones

objetivo a optimizar o implementar, así como también las necesidades de preguntas a resolver de parte de los datos recolectados".

(Benitez y Diez, 2005). "Uno de los pasos en los algoritmos de agrupamiento es el de asignar a cada objeto una medida de semejanza al patrón o centroide de cada cluster, con el fin de determinar a cuál de los grupos detectados pertenece el objeto en cuestión. Esta medida de semejanza entre objetos de un conjunto de datos se basa normalmente en el cálculo de una función de distancia. Seguidamente se exponen las más comunes, tanto para datos cuantitativos como cualitativos".

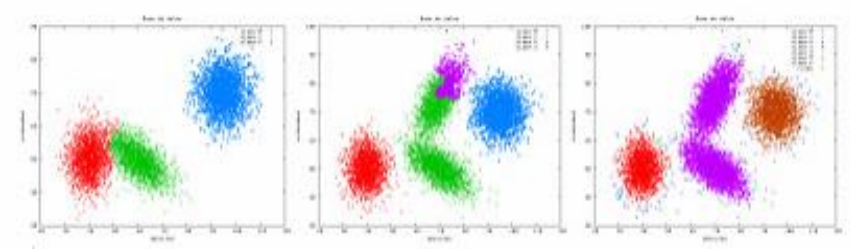


Figura 8 Datos agrupados

Fuente: Editorial Servicios de Publicaciones Universidad de Oviedo Introducción a la ingeniería Artificial: Sistemas Expertos, Redes Neuronales Artificiales y Computación Evolutiva.

2.3.3. Técnicas de agrupamiento para datos cualitativos y cuantitativos

Datos cuantitativos: los que son por position, por Span y por content, basadas en una modificación de distancia que podemos apreciar en la Figura 9.

$$
\begin{aligned}
d_p(A_k, B_k) &= \frac{|(al + a\mu)/2 - (bl + b\mu)/2|}{U_k} \\
d_s(A_k, B_k) &= \frac{|l_a - l_b|}{U_k + l_a + l_b - inters} \\
d_c(A_k, B_k) &= \frac{|l_a + l_b - 2inters|}{U_k + l_a + l_b - inters}
\end{aligned}
$$

Figura 9 Formula agrupación de medición cuántica por Span y Content.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

al = límite inferior de A_k
bl =límite inferior de B_k
$a\mu$ = límite superior de A_k
$b\mu$ = límite superior de B_k
$inters$ = longitud de la intersección entre A_k y B_k
$ls = span = |max(a\mu, b\mu) - min(al, bl)|$
U_k = diferencia entre el mayor y menor valor de la característica kth para todos los objetos
$l_a = |a\mu - al|$
$l_b = |b\mu - bl|$

Figura 10 Especificación de variables Span y Content.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

Datos cualitativos: solo se compone de dos dimensiones de similitud Span y Content, también basadas en distancias de Gowda y Diday, se muestra en la Figura 11.

$$D_s(A_k, B_k) = \frac{|l_a - l_b|}{l_s}$$
$$D_c(A_k, B_k) = \frac{|l_a + l_b - 2inters|}{l_s}$$

Figura 11 Formula cualitativa basada en distancias de centroides de Gowda y Diday.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

$\bar{inters}$ = número de elementos en la intersección entre A_k y B_k
ls = número de elementos en la unión entre A_k y B_k
l_a = número de elementos en A_k
l_b = número de elementos en B_k

Figura 12 Especificación de las variables de la fórmula de Gowda y Diday.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

La resultante: Algoritmos de clustering para datos cualitativos de funciones de pertenencia borrosa del algoritmo MVFCM. La grafica de muestra en la Figura 13 y el listado de Algoritmos de agrupamiento en la Figura 14.

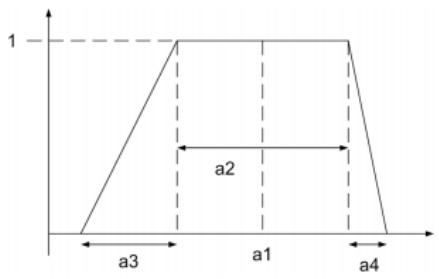


Figura 13 Algoritmo de clustering para datos cualitativos.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

	Clasificación	Función de similitud	Tipo de datos
K-modes	Particional	*Overlap metric*	Cualitativos
K-prototypes	Particional	*Overlap metric*	Cualitativos y cuantitativos
COOLCAT	Particional	Entropía	Cualitativos
ROCK	Jerárquico	Enlaces o vecinos comunes	Cualitativos
LIMBO	Jerárquico	Dist. de Kullback-Leibler	Cualitativos
STIRR	Basado en grafos	Co-ocurrencias	Cualitativos
CACTUS	Basado en grafos	Conexiones fuertes	Cualitativos
CLICKS	Basado en grafos	Densidad	Cualitativos
KEROUAC	Datos distribuidos	*New Condorcet Criterion* (NCC)	Cualitativos
MVFCM	Borroso	Dist. de Gowda y Diday modificada, y dist. según los TFN	Cualitativos, cuantitativos y funciones de pertenencia borrosas

Figura 14 Lista de algoritmos de Clasificación de agrupamiento algoritmos.

Fuente: Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

2.4. Minería de datos de datos

Las bases de datos han ido creciendo y la información ha ido en aumento en los últimos años en la cual se vive con la generación millennials, los cuales generan a cada segunda información de distinta naturaleza: numérica, selección categórica, etc. Las cuales se convierten en muchos campos y variables de distinta naturaleza, acotados a rangos de posibles valores. Representando una información de refinamiento, extra o de concreción, de los cuales hay campos en blanco o información errónea, el cual origina ruido para el análisis posterior. La necesidad de extraer conocimiento útil de la gran cantidad de datos el cual permita establecer relaciones entre el conjunto de datos con tal de simplificar la vista en la cual apliquemos técnicas de clasificación, agrupamiento, etc.

2.4.1. Base de datos para la minería de datos

Las bases de datos tienen muchos objetos de los cuales están representados por un conjunto de atributos, dimensiones o características, normalmente representadas de forma vectorial. Los datos son una serie de hecho.

según Jain y Dubes (1988). “la clasificación de datos se realiza en tipos y en escalas. El tipo de datos se refiere a su grado de cuantificación, es decir, que rango de valores pueden

abarcar y si estos son continuos o discretos. Una característica es continua si existen infinito número de valores posibles entre dos valores cualesquiera que pueda tomar la característica. Por ejemplo, mediciones de sensores o el volumen de un objeto son variables consideradas continuas. Por el contrario, una característica es discreta si todos los elementos del dominio al que pertenece pueden enumerarse en una correspondencia de un subconjunto finito de enteros positivos. Por ejemplo, la edad, el número de hijos, o los números ordinales se consideran características discretas. Unos casos especiales de atributos discretos son aquellos que solo pueden tomar dos valores, como por ejemplo las respuestas con solo dos posibilidades (Sı – No, 1 – 0). Estas características reciben el nombre de características binarias".

2.5. Minería de datos

(Weiss y indurkhya, 1998). "La minería de datos ha dado lugar a la sustitución del análisis de datos La principal diferencia entre ambos se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis. La automatización de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad. Dichas técnicas reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones".

según (Weiss y indurkhya, 1998). "Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento".

Supervisados o predictivos: Predicen el valor de un atributo de conjunto de datos. Weiss y indurkhya (1998) “De estos datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos)”.

El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. En la Figura 15 se muestran algunas de las técnicas de minería de ambas categorías.

Figura 15 Categorías de algoritmos de minería de datos.

Fuente: Publicado por Moisés perla - Gestión de negocios con data Warehouse y Data Mining.

Las técnicas de minería de datos y modelos predictivos.

Relevancia de la preparación de datos:

- Tener como resultado a modelos poco útiles, ya que los datos pueden estar incompletos, datos con ruido o inconsistentes.
- La preparación de los datos puede reducirlos con las técnicas de limpieza de datos ya sea por la eliminación de registros duplicados, anomalías, etc.

➢ la preparación tiene datos de calidad y tener modelos de calidad.

El proceso de extracción de conocimiento o *Knowledge Discovery in Databases*, es el proceso para identificar patrones que puedan ser útiles de algún modo dando información de millones de datos utilizando otras técnicas que no sean el de las estadísticas. Se utiliza en rubros como la inteligencia artificial, sistemas de gestión de bases de datos, sistemas de apoyo a la toma de decisiones en diferentes áreas y las etapas del proceso que están en la Tabla 1, los procesos de data mining en la figura 16 y la clasificación de Data Mining en la Figura 17.

Tabla 1

Etapas del proceso de Data Mining.

Selección de datos	Es la etapa inicial, es dónde se define qué datos serán recolectados, qué tipo de extracción tendrán, qué atributos de entrada y salida habrá, la justificación sobre por qué obtener los datos que se pretende conseguir, junto con las fuentes que puedan ser útiles.
Data Warehouse	Se diseña el esquema de un almacén de datos que consiga unificar de manera eficiente toda la información recogida.
Implantación del almacén de datos	Se instala la estructura o sistema que permita navegar entre los datos y así discernir qué información puede ser utilizada para analizar a profundidad.
Limpieza de datos	Se seleccionan, limpian y transforman los datos que se analizarán.
Selección de técnica	Teniendo los datos ya limpios se selecciona la técnica de minería de datos más apropiada para el fin que ya se definió en el primer paso.
Interpretación	Se evalúan diferentes aspectos de los datos procesados; coherencia, apego a la realidad, utilidad, aplicación en casos hipotéticos, etc. Teniendo los datos ya procesados junto con las evaluaciones correctas, se "traducen" a los términos contextuales correspondientes al proceso y se extrapolan a los casos que ya se tengan contemplados.
Difusión	Se dan a conocer los resultados y se ponen en práctica.
	Diagrama de los pasos en el proceso KDD, anteriormente descritas.

Fuente: Minería de datos: conceptos y tendencias, 2016.

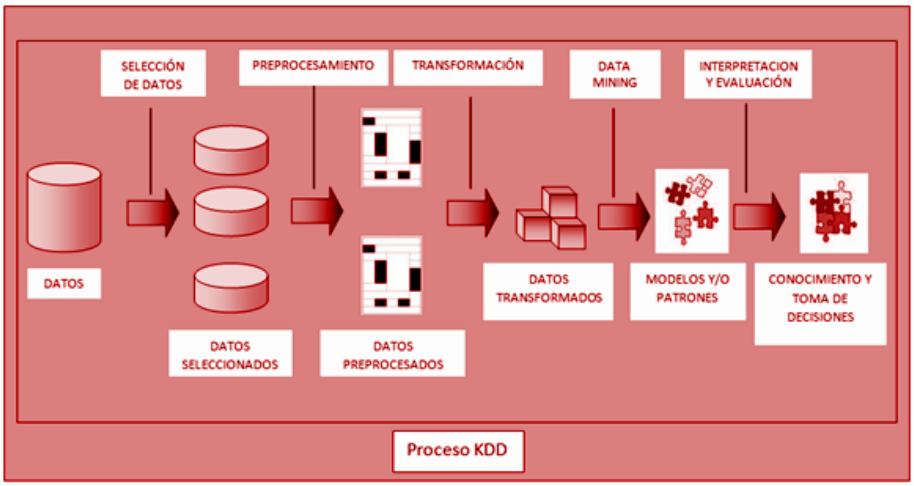


Figura 16 Procesos de KDD de minería de datos.

Fuente: Extraída de https://datosmineriainformacion.files.wordpress.com/2017/05/proceso_kdd_etapas.png

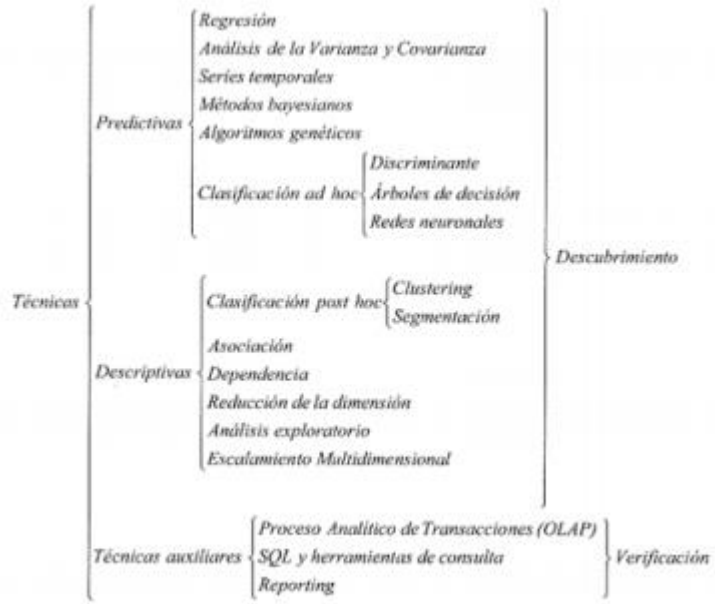


Figura 17 Clasificación de las técnicas de Data Mining
Fuente: Pèrez y santy, 2008.

2.6. Web scraping y minería de datos

2.6.1. Maquetación web

La extracción de datos de los sitios web de manera automática se refiere al proceso a la limpieza y filtrado de datos. que nos permite extraer datos escondidos de un documento ya sea páginas web o pdf, para que en el proceso hacerlos útiles y/o de estudio.

En esta sección es importante conocer la estructura de composición el HTML y/o de la página web ya que es importante para la extracción y utilización de los datos ya sea cualquier el giro de negocio y/o interés.

2.6.2. Extracción de datos

Habiéndonos centrado en que ítems de los que queremos extraer la información tenemos dos formas de hacerlo: mediante un API (Facebook, Twitter etc.) o web scraping (si una página no posee api esta es la mejor forma de consumir la información), las ventajas de utilizar web scraping es de no depender de una API (coste, licencia, numero de búsqueda, limitación de datos y consultas etc.), no se tiene limitaciones en cuanto a tiempo y qué información queremos obtener, lo que más debemos cuidar en el web scraping es la estructura de la página.

Tabla 2
Estructura de web scraping

Proceso del Web Scraping	
URL	Semilla
Request	Realizar requerimientos
Response	Obtener respuesta
Populate Items	Obtener la información que deseo de la respuesta. estructura
Más URLs	Ir a más URLS direcciones web y repetir el mismo proceso.

Fuente: Extraído de https://www.academia.edu/35895308/Web_scraping, 2018.

2.6.3. Tipos de web scraping

Una sola página web: las librerías que se pueden utilizar son *Scrapy.spiders.Spider*, solo tiene una página de extracción.

varias páginas web: Las librerías a utilizar *Scrapy.spiders.CrawlSpider:* Crawling vertical, Crawling horizontal, de los cuales se relaciona a la forma de separación de páginas que son en varias páginas de información dependiendo si son horizontales la paginación y vertical para los ítems de sacado de información más detallada.

Debemos de tomar en cuenta que hacer web scraping es darle crédito a la página de donde obtuvimos la Data, la no publicación de la data sin estar o comprobar que es seguro y/o legal, no sobrecargar las páginas ya que tendremos problemas con la IP, con web scraping somos rastreados y no somos anónimos.

2.6.4. Tipos de datos

En este punto la recolección de datos que tendremos que hacer son de datos estructurados y no estructurados, de los cuales haremos una diferenciación.

Tabla 3

Diferenciación de los tipos de datos

Datos estructurados	**Datos no estructurados**
Es aquella información que se encuentra en el modelo de la base de datos, las que se encuentran almacenadas en base de datos. Ya que pueden ser ordenados y procesados con la facilidad para el proceso de estudio de datos de las técnicas de minería de datos	Son las que se encuentran en forma de texto no almacenado y categorizado en un modelo de base de datos, Ya que no tienen infraestructura interna identificable, que es masivo, desordenado etc que produce ruido a la hora del análisis de datos.

Fuente: Extraído de https://www.academia.edu/35895308/Web_scraping, 2018.

2.7. Estudio de datos

2.7.1. Python para ciencia de datos

Por lo que sabemos Python es uno de los lenguajes más fáciles de entender y utilizar en donde nosotros buscamos utilizar para la manejabilidad de datos y así poder visualizar mediante sus librerías.

Librerías de visualización de datos con Python:

- Matplotlib: pandas seaborn - solo ideas de datos- estilos predeterminados no atractivo

- seaborn: se integra muy bien con pandas y otras bibliotecas de software de código abierto para análisis y visualización de datos. Es una librería popular para hacer atractivos gráficos de datos estadísticos en Python. Aprovecha el poder matplotlib para crear los gráficos con unas pocas líneas de código. (Estética, funciones, comparaciones. herramientas de visualización de modelos de regresión lineal para diferentes tipos de variables independiente o dependientes, funciones de visualizar matrices, abstracciones de alto nivel para estructurar grillas de parcelas)
- BOKEH: integrada con navegadores, javascript gráficos interactivos, tres interfaces: nivel alto, medio matplotlib, el más bajo para programadores.
- PYGAL: opciones de personalización, es de código abierto, capacidad de gráficos escalables.
- PLOTLY: Análisis de datos gráficos científicos. diferente a las demás para estar en línea.

Utilidad de librerías:

Tabla 4

Librerías de Python

	Descripción
MATPLOTLIB	Es el método más simple para las representaciones básicas
SEABORN	Es el ideal para crear gráficos estadísticos visualmente atractivos que incluyen color
BOKEH	Funciona muy bien para visualizaciones más complicadas e ideal para presentaciones interactivas basadas en web.
PYGAL	Funciona bien para generar vectores y archivos interactivos, sin embargo, no tiene flexibilidad como otros métodos.
PLOTLY	Es la opción más útil y fácil para crear visualizaciones altamente interactivas basadas en la web.

Fuente: Elaboración propia, 2019.

CAPÍTULO III: MATERIALES Y MÉTODOS

3.1. Descripción de lugar de ejecución

Los datos tomados a estudio con el algoritmo de clustering son datos de la Ciudad de Buenos Aires. Así como también datos para la elección de atributos en estudio fueron tomados de la policía nacional del Perú del área de turismo PENTUR - Puno, como también datos extraídos de la web de la ciudad de puno.

3.2. Metodología de la investigación

3.2.1. Tipo de investigación

El proyecto fue realizado con datos de delitos ciudadanos de Argentina, así como también un análisis de sistemas en la comisaría del PENTUR - Puno para la utilización de algoritmos, la extracción de datos a nivel de puno con las herramientas de web scraping. la investigación fue realizada de manera propositiva y aplicada.

3.2.2. Investigación propositiva

En el estudio de relación de factores fueron realizadas las actividades de:

- Técnica de web scraping: con spider y spider crawl para la extracción de datos de una sola página y de varias páginas utilizando xpath, Python y anaconda como herramientas tecnológicas para la obtención de datos de la región de Puno en cuanto al turismo.
- Análisis de sistemas: fue realizado en la Policía nacional del Perú realizando la base de datos de denuncias de delitos y faltas turísticas. este estudio fue con el fin de identificar los atributos.

3.2.3. Investigación aplicada

➢ Estudio de datos con los siguientes: Ciencia de datos, conocimientos de big data, Weka, ciencia de datos para Python y visualización de datos.

3.2.4. Arquitectura de solución

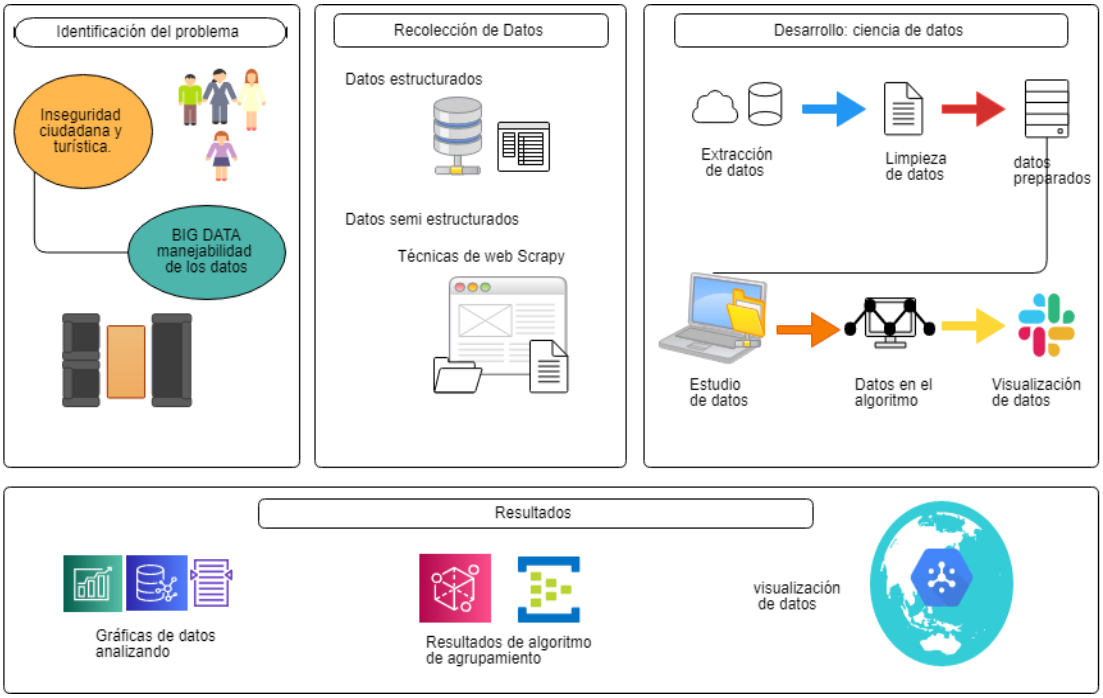


Figura 18 Arquitectura de solución al problema.

Fuente: Elaboración Propia, 2019.

3.3. Herramientas tecnológicas

El modelo de procesamiento del proyecto según su naturaleza y su aplicación, así como las herramientas a utilizar en el desarrollo para la extracción de datos de la Tabla 5.

Tabla 5
Herramientas de extracción de datos.

Herramienta	Descripción de la utilización
Python 3	Para la instalación de los manejadores pip y librerías. Que se utilizan en el desarrollo de la investigación.
Xpath	Para la maquetación y estructuración de la extracción de datos.
Scraping 3	Librería con la cual realizamos el scrapy, con las librerías principales de spider y spider scrawl.

IDE	Sublime text que es utilizado tanto para la extracción de datos como para la maquetación de la implementación del sistema de visualización como propuesta.

Fuente: Elaboración Propia, 2019.

Estudio de datos:

Tabla 6
Herramientas d estudio de datos.

Herramienta	Descripción de la utilización
WEKA	weka es una de las herramientas más completas para la minería de datos, que pueden generar modelo y patrones el cual construye modelos predictivos. En el proyecto es utilizado para el estudio de datos, reconocimiento de variables, y análisis de clustering con el algoritmo de Kmeans.
Python anaconda	Estudio de datos con ciencia de datos para Python. anaconda es una distribución de Python la ventaja de utilizar anaconda en que ya no se necesita instalar por separado las librerías de Jupyter notebook, pandas, librerías de visualización de datos de forma gráfica, etc. En el proyecto es utilizado para la limpieza de datos, para el análisis y graficación de datos.

Fuente: Elaboración Propia, 2019.

Kmeans

El K -means es un algoritmo de agrupamiento utiliza el refinamiento iterativo para producir un resultado final. Las entradas del algoritmo son el número de grupos K y el conjunto de datos. El conjunto de datos es una colección de características para cada punto de datos. El algoritmo se inicia con estimaciones iniciales para la kappa centroides, que o bien se pueden generar aleatoriamente o se selecciona entre el conjunto de datos al azar" (Trevino, 2016).

Visualización de datos:

Tabla 7
herramientas de visualización de datos

Herramientas	Descripción de la utilización
power bi	Power BI es un servicio de análisis de negocios de Microsoft. Su objetivo es proporcionar visualizaciones interactivas y

	capacidades de inteligencia empresarial con una interfaz.
	utilizado para la clasificación de los grupos de centroides y sus objetos
Python	Estudio de datos con ciencia de datos para Python. anaconda es una distribución de Python la ventaja de utilizar anaconda en que ya no se necesita instalar por separado las librerías de Jupyter notebook, pandas, librerías de visualización de datos de forma gráfica, etc. En el proyecto es utilizado para la limpieza de datos, para el análisis y graficación de datos.

Fuente: Elaboración Propia, 2019.

Estructuración de implementación de estudios de datos e ingeniería de software:

Tabla 8
herramientas de implementación de datos

Herramientas
Python - Django
Materialize
D3
JavaScript
sql

Fuente: Elaboración Propia, 2019.

3.4. METODOLOGÍA

La gestión del desarrollo del producto se realizará con la metodología OpenUP la cual es una metodología ágil para el desarrollo de proyectos ya que es flexible para el cambio y/o aumento de requerimientos, al tamaño pequeño del equipo de trabajo. La gestión del desarrollo del estudio de datos será con CRISP - DM.

3.4.1. Metodología para la gestión de proyectos

La metodología OpenUP es un Proceso unificado que aplica enfoques iterativos e incrementales dentro del desarrollo del proyecto, nos ofrece una naturaleza enfocada a la colaboración en el proceso de desarrollo de software, así como también un alto grado de adaptabilidad a las necesidades de un proyecto particular por su carácter iterativo. Esta

metodología está basada en RUP (Rational Unified Process). Lo que ofrece para el equipo de desarrollo esta metodología es: calidad y eficiencia.

(Rios, Hinojosa, Delgado, 2013). “Esta metodología fue propuesta por el grupo de empresas conformado por: IBM Corp, Telelogic AB, Armstrong Process Group Inc., Number Six Software Inc. y Xansa; quienes la donaron a la Fundación Eclipse en el año 2007, que la ha publicado bajo licencia libre”.

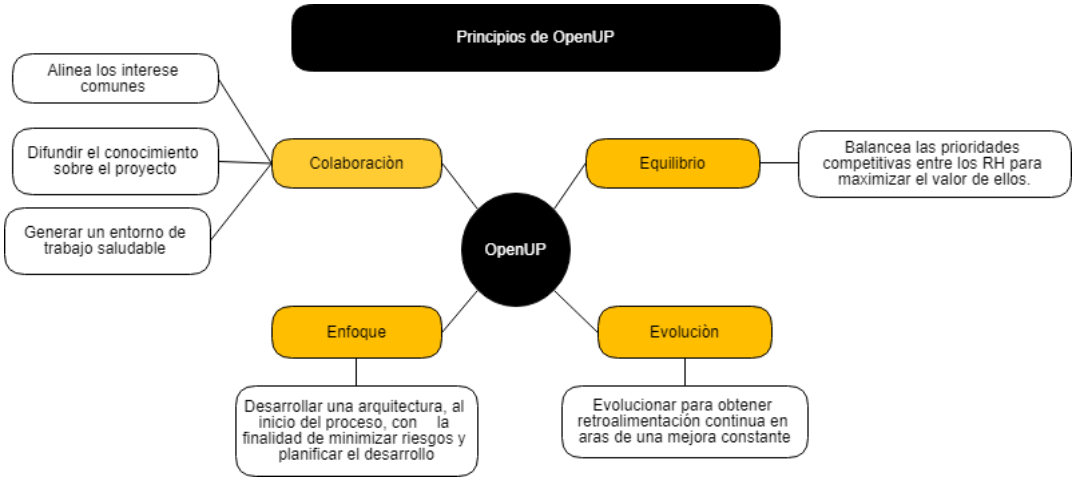


Figura 19 Principios de OpenUp

Fuente: Elaboración Propia.

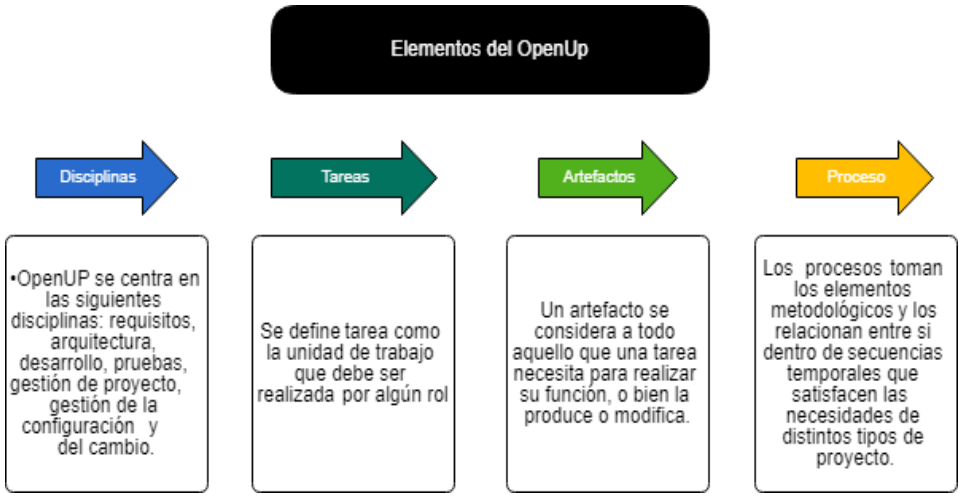


Figura 20 Elementos del OpenUp

Fuente: Elaboración propia

Ciclo de vida de OpenUP

La metodología de OpenUP consta de cuatro fases: inicio, elaboración, construcción y transición. Cada una de estas fases se divide a su vez en iteraciones.

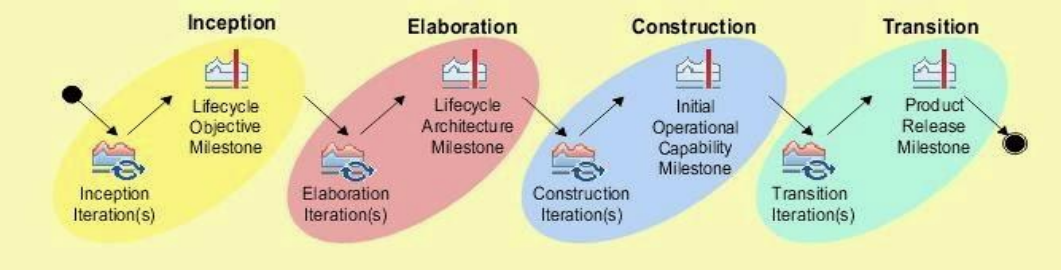


Figura 21 Ciclo de vida de OpenUP

Fuente: Néstor Díaz & David Vaquero-aplicación de la metodología openup en el desarrollo del sistema de difusión de gestión del conocimiento de la espe.

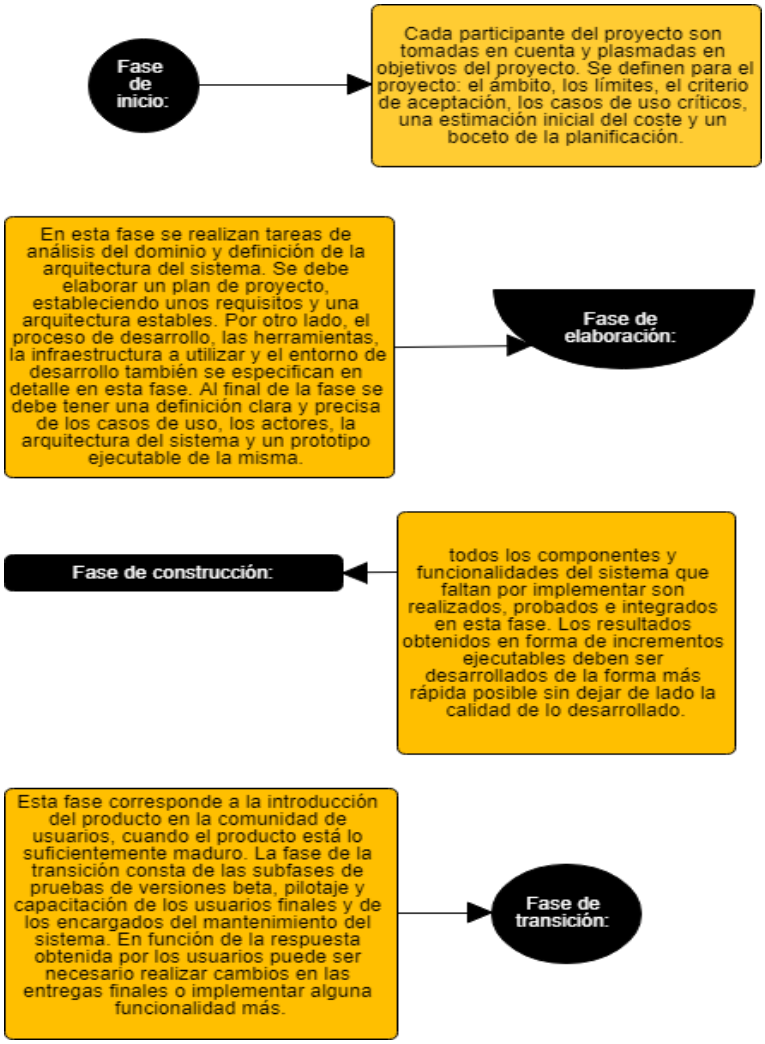


Figura 22 Fases de OpenUP

Fuente: Elaboración propia inspirado de Néstor Díaz & David Vaquero-aplicación de la metodología openup en el desarrollo del sistema de difusión de gestión del conocimiento de la espe.

3.4.2. Metodología de aplicación CRISP-DM

Rodriguez y Garcia, 2016). “Metodología CRISP-DM (Cross-Industry Standard Processfor Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos), es creada en el 2000 por el grupo de empresas SPSS, NCR y Daimer Chrysler. Es de distribución libre lo que le permite estar en constante desarrollo por la comunidad internacional. Además, resulta independiente de la herramienta que se utilice para llevar a cabo el proceso de MD. Es ampliamente usado por los miembros de la industria”.

(Chapman, Clinton, Kerber, Khabaza, reinartz, Shearer y Wirth, 2000) “El modelo consiste en seis fases definidas de manera cíclica: análisis del problema, comprensión de datos, preparación de datos, modelado, evaluación y despliegue”.

(Rodriguez y Garcia, 2016). “Debido a que la cantidad de datos almacenados, de todo tipo, van en aumento exponencial, existe la necesidad de tener mecanismos eficientes para manipularlos y extraer conocimientos de ellos. La minería de datos es de las principales encargadas de este tipo de proceso y para hacer menos complejos sus procedimientos se han diseñado metodologías que los guíen. Debido a que estas metodologías son de propósito general en ellas no se describen cuestiones importantes como técnicas y algoritmos a usar en cada etapa. En la presente investigación, luego de un estudio comparativo, se escoge la metodología CRISP-DM para realizar su adecuación a problemas no supervisados tipo atributo-valor. Los procesos de MD tienen muchas veces implícitas técnicas de aprendizaje que, de acuerdo con la definición dada por Michalski en 1986, es la habilidad de adquirir nuevo conocimiento, desarrollar habilidades para analizar y evaluar problemas mediante métodos y técnicas, así como también por medio de la experiencia propia; se requiere del aprendizaje entendible para un hombre”.

Fases de manera cíclica análisis del problema, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

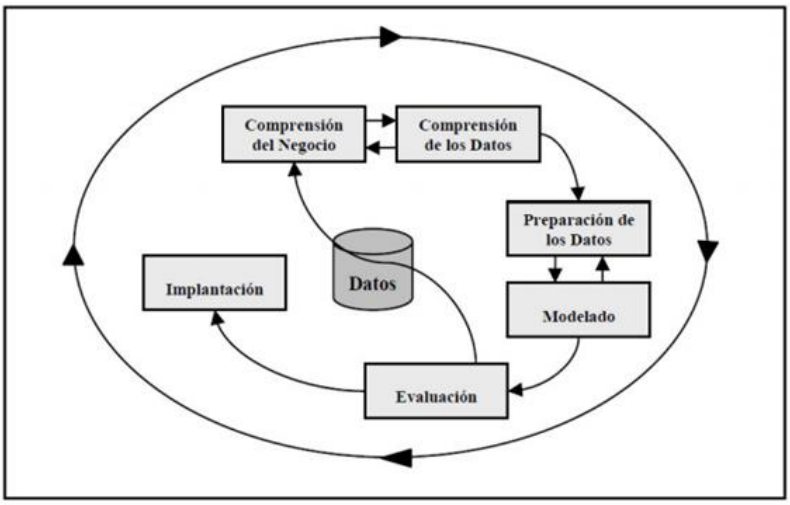


Figura 23 Modelo de proceso CRISP–DM ([CRISP-DM, 2000]).

Fuente: Cross-Industry Standard Processfor Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos.

Comprensión del negocio

- Entendimiento de los objetivos y requerimientos del proyecto.
- Definición del problema de Minería de Datos

Comprensión de los datos

- Obtención conjunto inicial de datos.
- Exploración del conjunto de datos.
- Identificar las características de calidad de los datos
- Identificar los resultados iniciales obvios.

Preparación de Datos

- Selección de datos
- Limpieza de datos

Modelamiento

- Implementación en herramientas de Minería de Datos

Evaluación

- Determinar si los resultados coinciden con los objetivos del negocio
- Identificar los temas de negocio que deberían haberse abordado

Despliegue

- Instalar los modelos resultantes en la práctica
- Configuración para minería de datos de forma repetida o continua.

3.4.3. Desarrollo de los objetivos

Para el logro de los objetivos de la investigación que es el análisis de datos en el cual en este capítulo se detalla la construcción de la minería de datos y los procesos para llegar al análisis de datos utilizando las herramientas de inteligencia artificial para llegar a la clustering de dichos datos y diseñar un prototipo implementando los algoritmos de clustering. en la cual se requiere acceder a los datos que permitan conocer el estado de delitos cometidos en contra de la ciudadanía para ponerlos en estadísticas en el contexto turístico.

3.4.4. Objetivo 01: **Recolectar datos estructurados y no estructurados con técnicas de web scraping.**

Para este objetivo se tomaron tanto datos de base de datos ya estudiadas como del mismo internet, que en breve explicaremos.

a. Datos estructurados

Los datos estructurados de la policía nacional del perú del área de turismo no se pudo recolectar ya que tiene políticas de privacidad de datos y gestión de permisos jerárquicos, se tramitó los documentos en el mes 20/07/2018 anexo N°5, pero no se encontró respuesta alguna por lo cual decidí trabajar con datos libres de la web **Data World** y los datos están publicados oficialmente en **mapa del delito** https://mapa.seguridadciudad.gob.ar/. Respecto a los datos de recolección del PENTUR recomendamos que las instituciones del estado

puedan liberar los datos para los estudios respectivos y no se haga tan complicada la obtención de datos ya que se perdió tiempo de estudio y proceso https://github.com/ramadis/delitos-caba/releases.

Los datos recolectados son dataset de los delitos ciudadanos de Argentina de la ciudad de Buenos Aires, del cual tenemos un archivo csv con 128 803 delitos registrados. los datos fueron tomados ya que se asemeja a los de las denuncias peruanas, con los mismos atributos poco cambiantes los cuales son:

Campo	Tipo	Descripción
id	Number	Identificador único del delito
comuna	String	Nombre de la comuna
barrio	String	Nombre del barrio
latitud	Number	Latitud del lugar del delito
longitud	Number	Longitud del lugar del delito
fecha	String	Fecha del delito en formato `YYYY-MM-DD`
hora	String	Hora del delito en formato `hh:mm:ss`
uso_arma	String	Describe si el robo fue armado
uso_moto	String	Describe si el robo fue por "motochorro"
origen_dato	String	Describe el origen de la denuncia
tipo_delito	String	Describe el tipo de delito
cantidad_vehiculos	Number	Cantidad de vehículos involucrados
cantidad_victimas	Number	Cantidad de víctimas fatales

Figura 24 Variables de datos Data Base

Fuentes: Base de datos de los datos de git hub, Buenos aires.

PENTUR **según análisis de sistemas** realizados en las prácticas pre profesionales adquirimos la forma de denuncia y también las variables de estudio que son los siguientes:

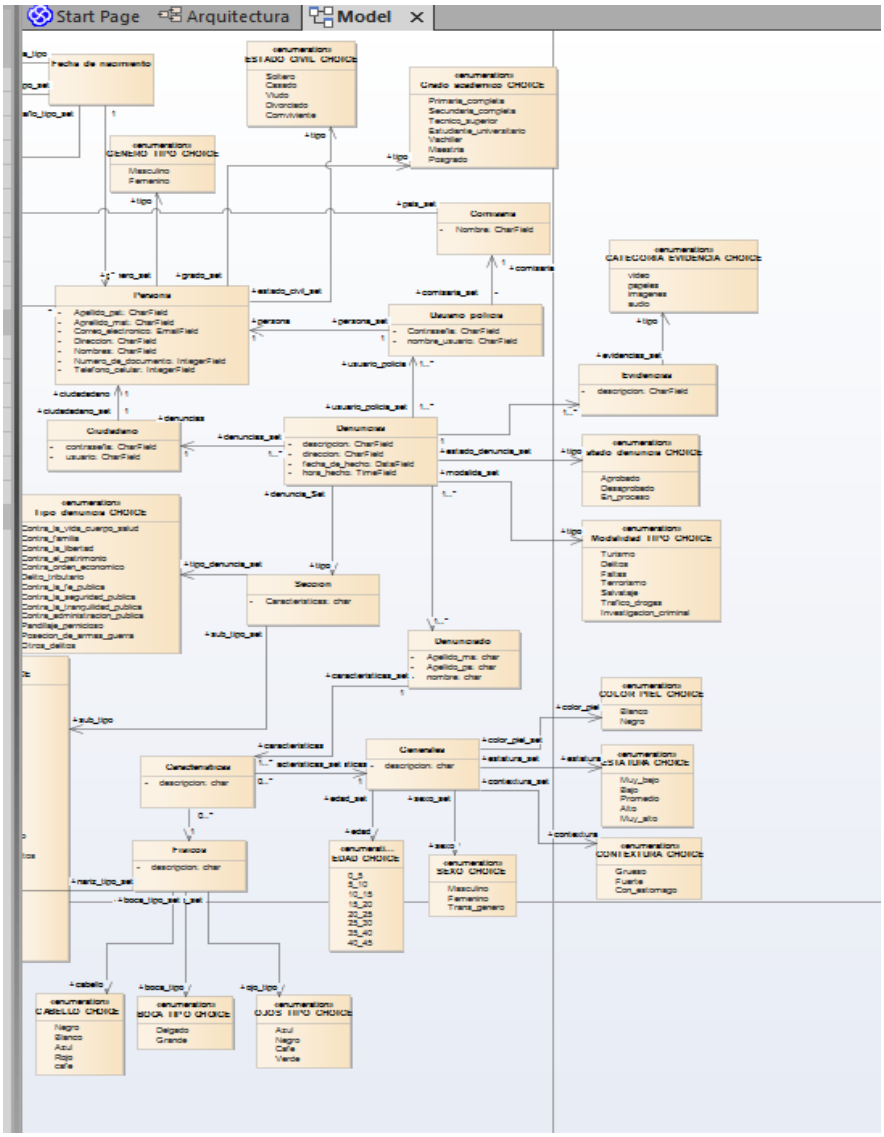


Figura 25 Base de datos

Fuente: Elaboración propia, 2018.

Los atributos seleccionados para trabajar son los siguientes:

Tabla 9
Variables para estudio

variable	tipo de variable
Departamento	string
recurso turístico / lugar turístico	string
latitud	int/float
longitud	int/float
fecha	date
hora	date
tipo de delito	string

Fuente: Elaboración propia, 2018.

Variables de estudio

El formato a tomar en cuenta para la variable de estudio es el: nombre del departamento, el recurso turístico o lugar turístico, latitud, longitud, fecha, hora, tipo de delito. de los datos recolectado de la ciudad de Buenos Aires  son los siguientes: comuna, barrio, latitud, longitud, fecha, hora, lugar, origen de dato, tipo delito.

b. Datos no estructurados

Los datos serán extraídos de la web utilizando técnicas de web scraping.

Web scraping

Los datos no estructurados los extraemos en esta ocasión de la web, serán datos que conforme la página web:

Tabla 10
Variables de extracción de datos.

Atributos	Variables
Nombre	TEXT
Calificaciones	TEXT
Colaboradores	TEXT
Número de fotos	TEXT
Dirección	TEXT
Ubicación	TEXT
Usuario	TEXT
Nombre del comentario	TEXT
Calificaciones	TEXT
Comentario	TEXT
Aceptación de comentario	TEXT

Fuente: Elaboración propia, 2018.

La página de la cual se extrajo los datos es de MI NUBE con 59 datos principales que son los recursos turísticos (https://www.minube.com/que_ver/peru/puno/puno).

Página de mi nube:

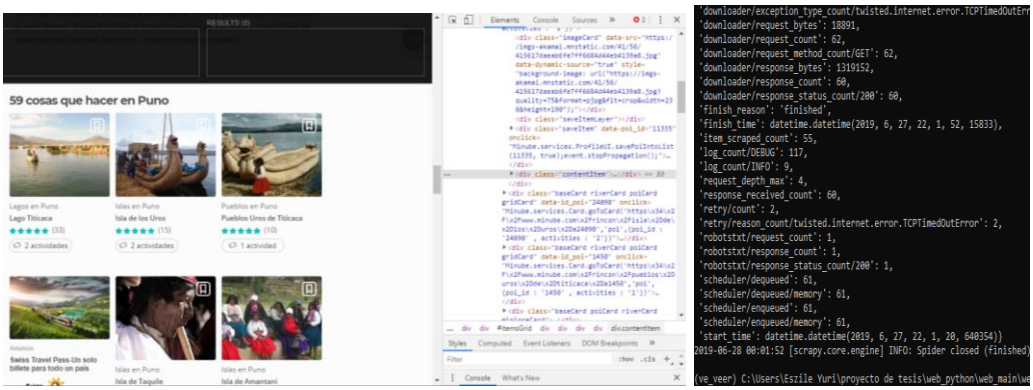


Figura 26 Página de extracción de datos

Fuente: Elaboración propia, 2019.

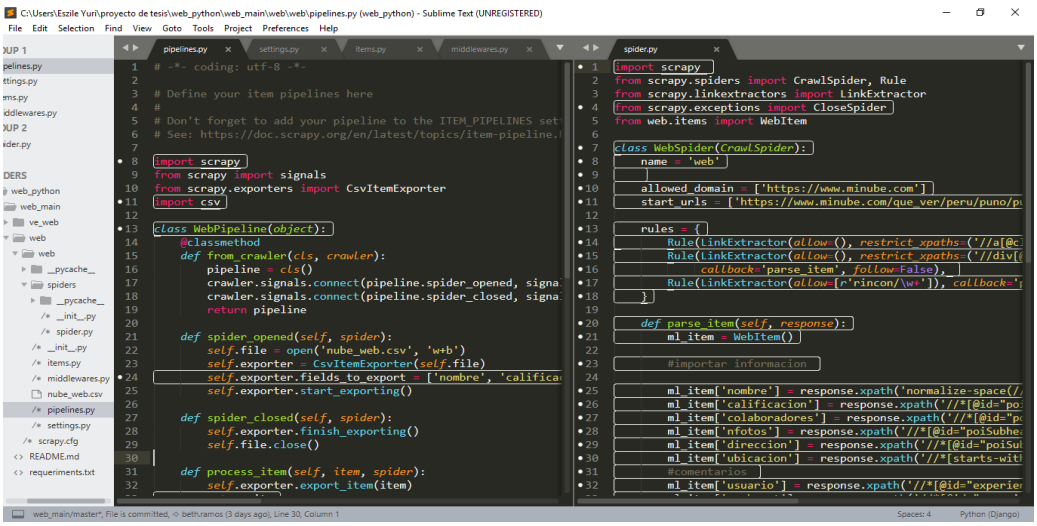


Figura 27 Código de extracción de datos.

Fuente: Elaboración propia, 2019.El código y los datos recolectados se encuentran en el repositorio de GIT LAB https://gitlab.com/beth.ramos/web_main.

3.4.5. Objetivo N° 2**: Analizar los datos recolectados con Python para ciencia de datos y Weka.**

Estudio de datos estructurados

El dataset utilizado para el tratamiento de datos es de crimen y delito registrado en la ciudad de buenos aires, dataset registra 184.877 delitos, registrados por el gobierno de la ciudad autónoma de buenos aires, durante el periodo de 07/11/2015 hasta el 30/06/2017 los datos fueron oficialmente publicados en la página de **mapa del delito** https://mapa.seguridadciudad.gob.ar/.

La descripción de datos recolectados tiene una gran similitud con la de la policía nacional del Perú indicando los campos de:

Tabla 11
Variables de estudio.

Campo	Tipo	Descripción
id	Number	Identificador único del delito
comuna	String	Nombre de la comuna
barrio	String	Nombre del barrio
latitud	Number	Latitud del lugar del delito
longitud	Number	Longitud del lugar del delito
fecha	String	Fecha del delito en formato YYYY-MM-DD
hora	String	Hora del delito en formato hh:mm:ss
uso_arma	String	Describe si el robo fue armado
uso_moto	String	Describe si el robo fue por "motochorro"
lugar	String	Describe el lugar del delito
origen_dato	String	Describe el origen de la denuncia
tipo_delito	String	Describe el tipo de delito
cantidad_vehiculos	Number	Cantidad de vehículos involucrados
cantidad_victimas	Number	Cantidad de víctimas fatales

Fuente: Elaboración Propia, 2018.

Para el trabajo en weka encontramos los siguientes datos generales de los 184.877 delitos:

Figura 28 . Weka y atributos de los datos

Fuente: Elaboración propia, 2019.

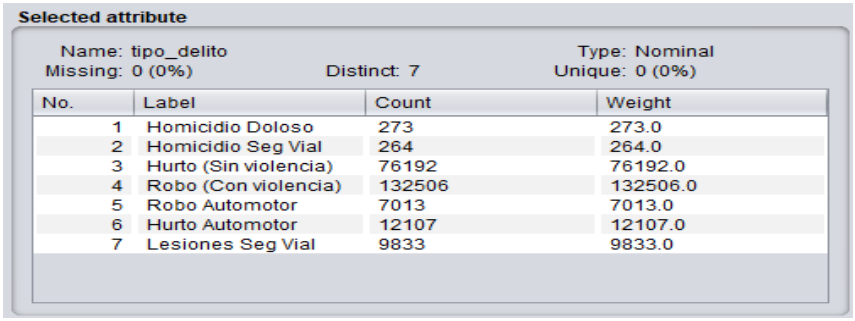


Figura 29 Weka y atributos de los datos

Fuente: Elaboración propia, 2019.

Tabla 12

Variables e identificación de atributos de los datos

Comuna son 15 como se muestra en la figura n° 22
barrio son 15 como se muestra en la figura n° 22
7 tipos de delitos como se muestra en la figura n° 23
tenemos datos de 2016 a 2017 tenemos 731 fechas y 1417 horas
sin uso de armas y asimismo sin uso de motos

Fuente: Elaboración Propia, 2018.

Limpieza de datos estructurados

Para hacer la corrección o eliminación de registro de datos de la base de datos identificamos datos incompletos haciendo una data duty, limpieza de datos:

Los datos limpiados son tanto para weka y Python anaconda.

Borramos dos atributos que no tenían valores almacenados como lo son lugar y **origen_dato** para no provocar algún ruido en los resultados de clustering.

los atributos de **uso_arma** y **uso_moto** también fueron borrados por que no varían en su valor que es *sin uso de arma* y *sin uso de moto*, por lo tanto, los delitos cometidos serían sin estos dos atributos.

Por último, también borramos **cantidad_vehículos** ya que no cuenta con valores superiores a 0.0 de los cuales podríamos decir que los delitos cometidos son sin vehículos.

Limpieza de datos weka:

Relation: data

No.	1: id	2: comuna	3: barrio	4: latitud	5: longitud	6: fecha	7: hora	8: uso_arma	9: uso_moto	10: lugar	11: origen_dato	12: tipo_delito	13: cantidad_vehiculos	14: cantidad_victimas
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	String	String	Nominal	Numeric	Numeric
1	684...	Comun...	BALV...	-34.6...	-58.4117	2016...	01:0...	SIN USO ...	SIN MOTO			Homicidio ...	0.0	0.0
2	684...	Comun...	FLO...	-34.6...	-58.4547	2016...	02:3...	SIN USO ...	SIN MOTO			Homicidio ...	0.0	0.0
3	684...	Comun...	NUE...	-34.6...	-58.4053	2016...	04:0...	SIN USO ...	SIN MOTO			Homicidio ...	0.0	0.0

Figura 30 data con atributos originales

Fuente: Elaboración propia, 2019.

Relation: data

No.	1: id	2: comuna	3: barrio	4: latitud	5: longitud	6: fecha	7: hora	8: uso_arma	9: uso_moto	10: tipo_delito	11: cantidad_victimas
	Numeric	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric
1	684...	Comun...	BALV...	-34.6...	-58.4117	2016...	01:0...	SIN USO ...	SIN MOTO	Homicidio ...	0.0
2	684...	Comun...	FLO...	-34.6...	-58.4547	2016...	02:3...	SIN USO ...	SIN MOTO	Homicidio ...	0.0
3	684...	Comun...	NUE...	-34.6...	-58.4053	2016...	04:0...	SIN USO ...	SIN MOTO	Homicidio ...	0.0

Figura 31 Data limpia

Fuente: Elaboración propia, 2019.

Limpieza de datos Python

En los datos semi estructurados no hicimos la limpieza de datos, ya que solo contamos con 59 datos de la página mi nube.

el desarrollo lo trabajamos con weka y python.

3.4.6. Objetivo N° 3**: Preparar los datos y aplicar algoritmos de clustering a los datos recolectados.**

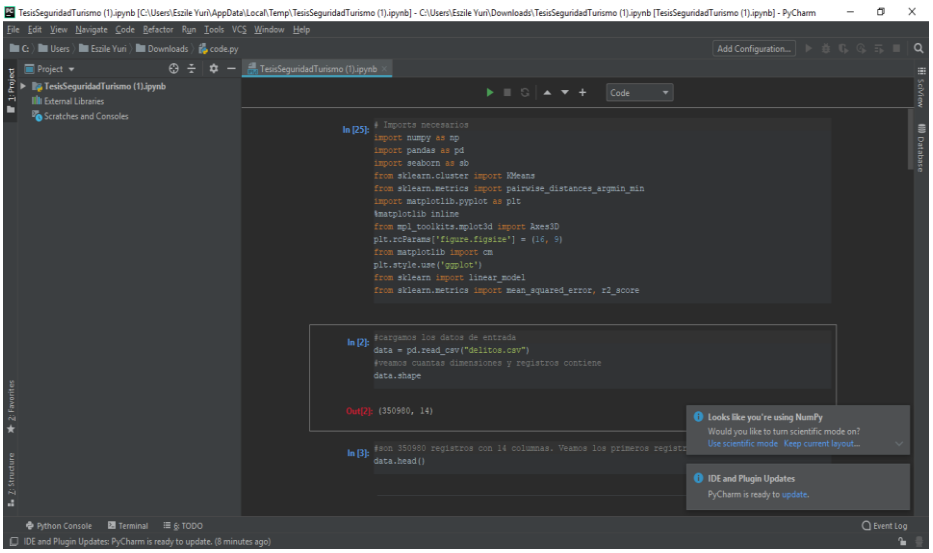


Figura 32 Entorno de desarrollo

Fuente: Elaboración propia, 2019.

El algoritmo de clusterización no supervisada agrupa basándose en sus características de los objetos de k grupos:

Debido a la utilización de los datos no clasificados, cualitativos se tomaron grupos cruzados para su tratamiento en el algoritmo de clustering.

inicializamos con 14 k grupos, que no todos están establecido como centroides en el espacio de los datos de Argentina.

```
In [3]: #son 350980 registros con 14 columnas. Veamos los primeros registros
        data.head()
```

Out[3]:

	id	comuna	barrio	latitud	longitud	fecha	hora	uso_arma	uso_moto	lugar	origen_dato	tipo_delito	cantidad_vehiculos	cantidad_victi
0	68313	Comuna 7	FLORES	-34.647598	-58.442600	2016-01-31	19:50:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Homicidio Doloso	0	
1	130524	Comuna 9	LINIERS	-34.639801	-58.529701	2016-01-31	12:00:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	
2	130582	Comuna 9	LINIERS	-34.638802	-58.527500	2016-01-31	08:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	
3	130586	Comuna 10	VILLA REAL	-34.618000	-58.527302	2016-01-31	20:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Hurto Automotor	0	
4	130615	Comuna 9	LINIERS	-34.645599	-58.526501	2016-01-31	20:24:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	

Figura 33 Datos a estudio con JUPYTER

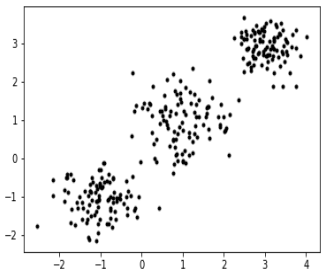
Fuente: Elaboración propia, 2019.

la asignación de los objetos de los centroides que son los más cercanos y posibles de trabajar son: tipos delitos, barrio, y comuna.

```
np.random.seed(7)

x1 = np.random.standard_normal((100,2))*0.6+np.ones((100,2))
x2 = np.random.standard_normal((100,2))*0.5-np.ones((100,2))
x3 = np.random.standard_normal((100,2))*0.4-2*np.ones((100,2))+5
X = np.concatenate((x1,x2,x3),axis=0)

plt.plot(X[:,0],X[:,1],'k.')
plt.show()
```



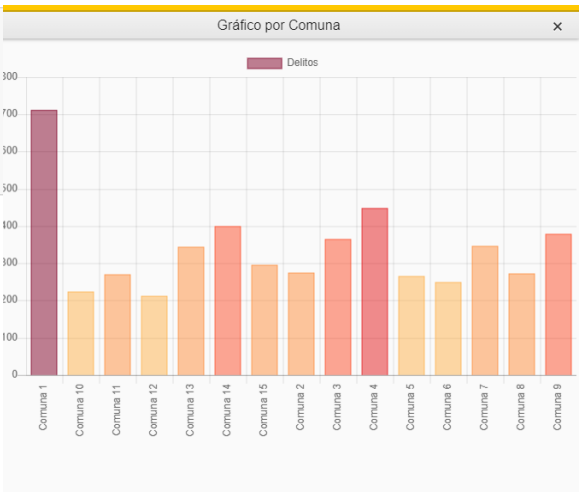


Figura 34 resultados de comuna.

Fuente: Elaboración propia, 2019.

El algoritmo de clustering resuelve los problemas de la suma de las distancias cuadráticas de cada objeto al centroide del objeto ya que es un problema de optimización.

$$\min_{\mathbf{S}} E(\boldsymbol{\mu_i}) = \min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \quad (1)$$

se toma el promedio de los elementos de cada grupo como un nuevo centroide.

$$\frac{\partial E}{\partial \boldsymbol{\mu}_i} = 0 \implies \boldsymbol{\mu}_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$$

Este algoritmo converge en un mínimo local ya que los resultados dependerán mucho de los centroides. Importancia de la visualización de datos es una de las formas de mostrar datos complejos en una forma gráfica y fácil de entender, las tramas y los gráficos pueden ser muy eficaces para transmitir una descripción clara de los datos, Pueden ser muy valiosas cuando se trata de respaldar cualquier recomendación.

3.4.7. Objetivo N° 4**: Diseñar la plataforma para la visualización de datos.**

La plataforma de visualización de datos está siendo construida de acuerdo al giro de negocio, ya que las organizaciones como la PNP y serenazgo tiene esa responsabilidad de los datos y su procedencia. Los módulos distribuidos para su construcción son:

Figura 35 Estructura de módulos

Fuente: Elaboración propia, 2019.

El prototipo para su construcción es el siguiente, es una propuesta para su análisis con los respectivos especialistas en el giro del negocio.

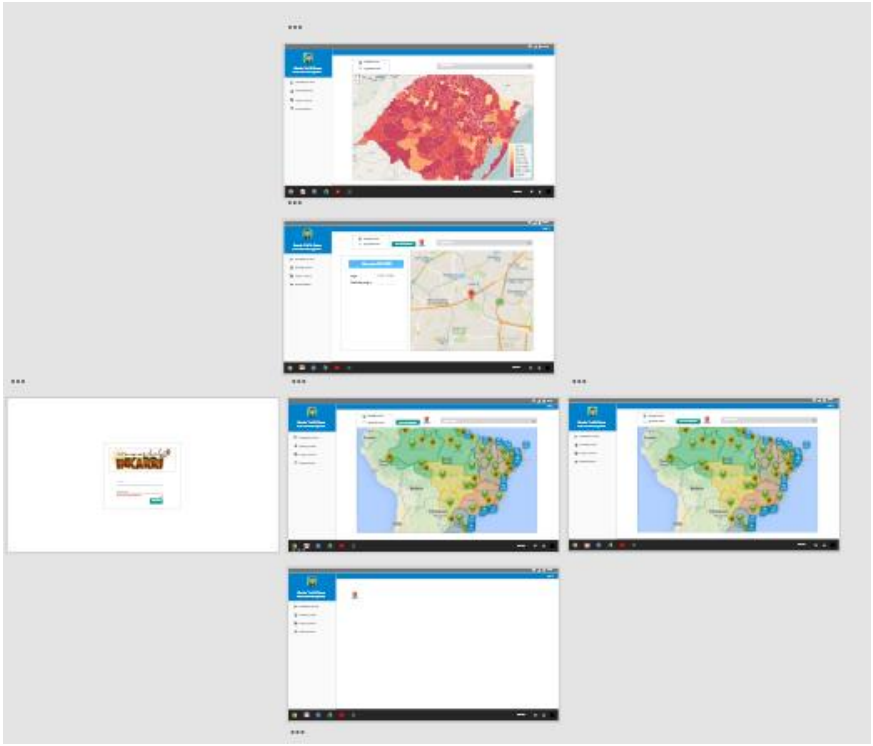


Figura 36 Prototipo de visualización.

Fuente: Elaboración propia, 2019.

las ventajas que se tiene de una buena visualización de datos con datos masivos, ya que forma un arte visual que capta el interés del analista de datos, los gráficos, es una tendencia de los valores atípicos, y sacar deducciones de historias a contar con un propósito rápidamente.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1. Resultados de los objetivos.

4.1.1. Objetivo 01: **Recolectar datos estructurados y no estructurados con técnicas de web scraping.**

Datos estructurados y semiestructurados.

El trabajo que se realizó en este objetivo fue la extracción de datos semi-estructurados de la web.

- Resultados: se sacaron 59 datos en los cuales están lo que buscamos sacar son los lugares turísticos para poder trabajar con eso y su aceptación, ya que para el turismo necesitamos como punto de centroide a los lugares turísticos para su clustering e puntos más cercanos de delitos.

  los datos extraídos son:

  Ese es el formato de extracción de los datos.

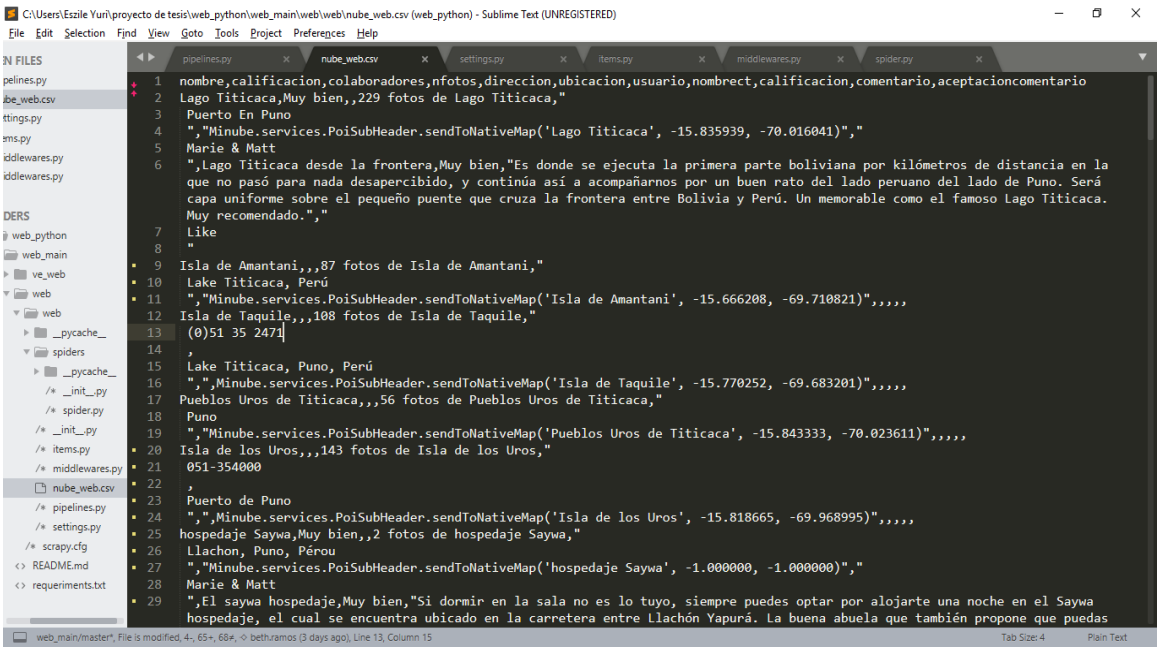


Figura 37 Datos extraídos de la página nube.

Fuente: Elaboración propia, 2019.

4.1.2. Objetivo N° 2**: Analizar los datos recolectados con Python para ciencia de datos y Weka.**

Análisis de weka por tipo delito y barrio.

```
Cluster 0: 'Comuna 9',2017-10-28,17:30:00,'Hurto (Sin violencia)',0,0
Cluster 1: 'Comuna 1',2017-05-15,00:00:00,'Robo (Con violencia)',0,0
Cluster 2: 'Comuna 8',2016-06-16,17:45:00,'Robo (Con violencia)',0,0
Cluster 3: 'Comuna 1',2017-05-04,20:20:00,'Robo (Con violencia)',0,0
Cluster 4: 'Comuna 5',2017-04-20,19:30:00,'Robo (Con violencia)',0,0
Cluster 5: 'Comuna 1',2016-04-19,00:00:00,'Robo Automotor',0,0
```

```
Cluster 0 <-- PALERMO
Cluster 1 <-- SAAVEDRA
Cluster 2 <-- VILLA LUGANO
Cluster 3 <-- SAN NICOLAS
Cluster 4 <-- ALMAGRO
Cluster 5 <-- FLORES
```

Attribute	Full Data	0	1	2	3
	(238188.0)	(62978.0)	(4841.0)	(9120.0)	(141036.0)
comuna	Comuna 1	Comuna 14	Comuna 1	Comuna 8	Comuna 1
fecha	2017-12-06	2017-10-28	2017-05-15	2016-06-16	2017-05-04
hora	00:00:00	20:00:00	00:00:00	17:45:00	21:00:00
tipo_delito	Robo (Con violencia)	Hurto (Sin violencia)	Robo (Con violencia)	Robo (Con violencia)	Robo (Con violencia)	Robo
cantidad_vehiculos	0	0	0	0	0
cantidad_victimas	0.0012	0.0003	0.0091	0.0195	0

Figura 38 Resultados de weka análisis de clustering.

Fuente: Elaboración propia, 2019.

De acuerdo al barrio con más incidencia delictiva.

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3
	(238188.0)	(73993.0)	(137292.0)	(3599.0)	(5322.0)
barrio	PALERMO	PALERMO	PALERMO	VILLA SOLDATI	CONSTITUCION
fecha	2017-12-06	2017-10-28	2017-05-15	2016-06-16	2017-05-04
hora	00:00:00	20:00:00	00:00:00	17:45:00	20:20:00
tipo_delito	Robo (Con violencia)	Hurto (Sin violencia)	Robo (Con violencia)	Robo (Con violencia)	Robo (Con violencia)	Robo
cantidad_vehiculos	0	0	0	0	0
cantidad_victimas	0.0012	0.0002	0.0001	0.0575	0.0028

Figura 39 Incidencias delictivas.

Fuente: Elaboración propia, 2019.

0	1	2	3	4	5	<-- assigned
7083	12202	80	56	270	230	| Comuna 3
4355	10739	68	39	274	2564	| Comuna 7
4628	13634	94	57	269	901	| Comuna 4
4565	8965	54	40	255	220	| Comuna 13
3967	9698	73	30	222	1034	| Comuna 9
4115	7625	42	27	171	344	| Comuna 11
3007	6039	2824	26	159	988	| Comuna 8
3946	8983	58	40	206	242	| Comuna 15
3153	7866	42	39	177	309	| Comuna 6
7177	11903	1	0	1	0	| Comuna 14
3627	2734	16	9	6819	350	| Comuna 5
3872	6861	40	23	127	55	| Comuna 2
13197	14302	100	4882	358	166	| Comuna 1
3384	8103	55	30	160	357	| Comuna 12
3917	7638	52	24	159	595	| Comuna 10

Figura 40 Centroides de incidencias.

Fuente: Elaboración propia, 2019.

Análisis por tipo delito.

Cluster 0 <-- Hurto (Sin violencia)
Cluster 1 <-- Robo (Con violencia)
Cluster 2 <-- Hurto Automotor
Cluster 3 <-- Homicidio Doloso
Cluster 4 <-- Lesiones Seg Vial
Cluster 5 <-- Robo Automotor

Class attribute: tipo_delito
Classes to Clusters:

0	1	2	3	4	5	<-- assigned to cluster
13	104	36	25	8	87	| Homicidio Doloso
29	162	36	15	9	13	| Homicidio Seg Vial
18587	8962	8123	13068	9335	18117	| Hurto (Sin violencia)
8839	46180	14564	21221	9632	32070	| Robo (Con violencia)
1038	2105	1388	809	531	1142	| Robo Automotor
1378	74	5650	2005	1037	1963	| Hurto Automotor
776	6719	472	300	639	927	| Lesiones Seg Vial

Figura 41 Tipos delitos clasificados.

Fuente: Elaboración propia, 2019.

Attribute	Full Data	0	1	2	3	4	5
	(238188.0)	(30660.0)	(64306.0)	(30269.0)	(37443.0)	(21191.0)	(54319.0)
id	125685.4369	108509.3082	60861.1062	125797.1573	168509.7119	122293.5554	183864.7093
comuna	Comuna 1	Comuna 9	Comuna 1	Comuna 8	Comuna 1	Comuna 5	Comuna 3
barrio	PALERMO	MATADEROS	SAN NICOLAS	VILLA LUGANO	CONSTITUCION	ALMAGRO	BALVANERA
latitud	-34.5047	-34.4315	-34.4903	-34.4732	-34.5781	-34.593	-34.4958
longitud	-58.2546	-58.1472	-58.2322	-58.2015	-58.3779	-58.4029	-58.2287
fecha	2017-12-06	2017-10-28	2017-05-15	2016-06-16	2017-05-04	2017-04-20	2016-04-19
hora	00:00:00	17:30:00	00:00:00	20:00:00	20:00:00	19:30:00	00:00:00
uso_arma	SIN USO DE ARMA	SIN USO DE ARMA	SIN USO DE ARMA	SIN USO DE ARMA	SIN USO DE ARMA	SIN USO DE ARMA	SIN USO DE ARMA
uso_moto	SIN MOTO	SIN MOTO	SIN MOTO	SIN MOTO	SIN MOTO	SIN MOTO	SIN MOTO
cantidad_vehiculos	0	0	0	0	0	0	0
cantidad_victimas	0.0012	0.001	0.0027	0.0012	0.0004	0.0005	0.0002

Figura 42 Análisis de los cluster de tipo delito.

Fuente: Elaboración propia, 2019.

Los datos de análisis para la toma de centroides son los siguientes en ambas herramientas tecnológicas de análisis weka y Python: comuna, barrio, tipo delito, fechas.

Análisis de datos con Python:

a. Porcentaje de delitos en cada una de las comunas.

```
In [2]: #cargamos los datos de entrada
        data = pd.read_csv("delitos.csv")
        #veamos cuantas dimensiones y registros contiene
        data.shape

Out[2]: (350980, 14)
```

```
In [3]: #son 350980 registros con 14 columnas. Veamos los primeros registros
        data.head()
```

Out[3]:

	id	comuna	barrio	latitud	longitud	fecha	hora	uso_arma	uso_moto	lugar	origen_dato	tipo_delito	cantidad_vehiculos	cantidad_victi
0	68313	Comuna 7	FLORES	-34.647598	-58.442600	2016-01-31	19:50:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Homicidio Doloso	0	
1	130524	Comuna 9	LINIERS	-34.639801	-58.529701	2016-01-31	12:00:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	
2	130582	Comuna 9	LINIERS	-34.638802	-58.527500	2016-01-31	08:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	
3	130586	Comuna 10	VILLA REAL	-34.618000	-58.527302	2016-01-31	20:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Hurto Automotor	0	
4	130615	Comuna 9	LINIERS	-34.645599	-58.526501	2016-01-31	20:24:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	

Figura 43 Resultados de comuna en Python.

Fuente: Elaboración propia, 2019.

```
In [4]: # Ahora veamos algunas estadísticas de nuestros datos
        data.describe()
```

Out[4]:

	id	latitud	longitud	lugar	origen_dato	cantidad_vehiculos	cantidad_victimas
count	350980.000000	350980.000000	350980.000000	0.0	0.0	350980.0	350980.000000
mean	183126.690207	-34.233231	-57.796189	NaN	NaN	0.0	0.001174
std	104320.063799	3.602982	6.082883	NaN	NaN	0.0	0.035708
min	1.000000	-34.849583	-58.958470	NaN	NaN	0.0	0.000000
25%	97578.750000	-34.631699	-58.469381	NaN	NaN	0.0	0.000000
50%	185323.500000	-34.610298	-58.433374	NaN	NaN	0.0	0.000000
75%	273079.250000	-34.592526	-58.400101	NaN	NaN	0.0	0.000000
max	360824.000000	0.000000	0.000000	NaN	NaN	0.0	3.000000

Figura 44 Comuna resultados de combinaciones.

Fuente: Elaboración propia, 2019.

```
In [5]: sb.factorplot('comuna',data=data,kind="count",aspect=3)

Out[5]: <seaborn.axisgrid.FacetGrid at 0x1e5e1471da0>
```

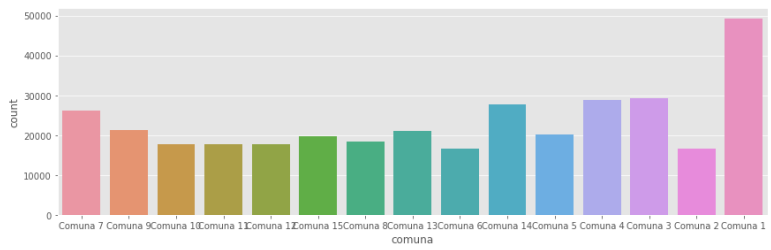


Figura 45 Comuna gráfica de barras

Fuente: Elaboración propia, 2019.

La comuna es la que tiene el más alto grado de interacción de delitos, el cual tiene el más alto nivel de delincuencia, en su totalidad 50000, seguido de la comuna 3, comuna 4, la comuna 7 y la comuna 14 son los que tienen más alto porcentaje de actividad delictiva.

b. Análisis de tipo de delito y constancia de valor mayor de sucesos.

los datos son semejantes a lo de la policía nacional del Perú en cuanto a los tipos de delitos.

Tabla 13
Delitos y faltas.

Delitos	Faltas
Hurto (apropiarse de un bien ajeno sin violencia).	Hurto (apropiarse de un bien ajeno sin violencia).
Robo (Acto de apoderarse de un bien ajeno, en la cual hay violencia, amenaza o fuerza.)	Participación en juegos inapropiados
Estafa	Otros
Apropiación ilícita (provecho o en el de un tercero; haciendo suya en forma indebida un bien mueble, una suma de dinero o cualquier objeto que se haya entregado para la guarda o depósito)	
Otros delitos contra el patrimonio otros delitos	

Fuente: Elaboración Propia, 2018.

Como podemos ver en las gráficas los datos muestran que el 57.9 por ciento son de un suceso con el delito de robo con violencia, seguido con el de hurto sin violencia.

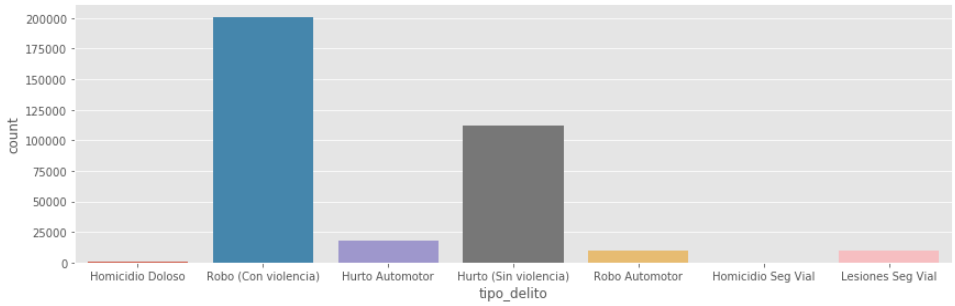


Figura 46 hurto sin violencia. gráfica de barras.

Fuente: Elaboración propia, 2019

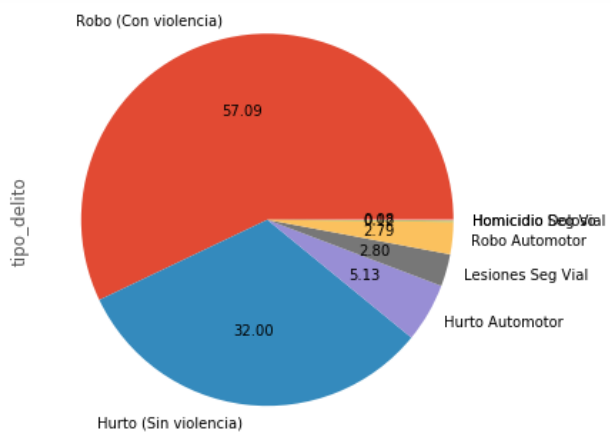


Figura 47 hurto sin violencia gráfica

Fuente: Elaboración propia, 2019

tipo_delito comuna	Homicidio Doloso	Homicidio Seg Vial	Hurto (Sin violencia)	Hurto Automotor	Lesiones Seg Vial	Robo (Con violencia)	Robo Automotor
Comuna 1	21.481481	14.772727	18.272282	3.086177	12.722465	13.349936	3.746012
Comuna 10	2.962963	3.409091	5.013259	8.387206	5.166277	4.662793	8.088916
Comuna 11	2.716049	4.924242	5.226302	6.464669	5.664599	4.846427	5.474941
Comuna 12	2.962963	6.818182	5.216414	7.684524	6.203600	4.746812	4.425234
Comuna 13	1.481481	4.545455	6.332869	6.981843	5.796807	5.976404	2.634558
Comuna 14	2.222222	6.060606	9.531215	5.621451	7.373131	7.629613	1.770094
Comuna 15	2.469136	4.166667	5.272147	6.414076	7.047697	5.890373	2.881548
Comuna 2	1.481481	2.272727	5.317992	1.821350	3.986576	4.977738	0.720387
Comuna 3	7.407407	4.545455	9.476381	4.182360	9.569816	8.346037	3.344654
Comuna 4	14.320988	10.227273	6.183649	8.128619	8.868097	9.174653	12.874344
Comuna 5	1.481481	2.651515	4.872129	7.459666	5.715448	6.138908	6.781929
Comuna 6	2.469136	3.787879	4.179963	5.969981	4.301841	5.082384	4.610476
Comuna 7	16.296296	9.469697	5.995775	8.016190	6.061222	8.038639	15.354533
Comuna 8	13.333333	12.500000	3.902198	9.185452	4.128954	5.418459	12.761140
Comuna 9	6.913580	9.848485	5.207425	10.596436	7.393471	5.720826	14.531234

Figura 48 hurto sin violencia R combinaciones.

Fuente: Elaboración propia, 2019

Análisis de acuerdo al barrio en donde podemos rescatar que el barrio con más índice de delitos es el de Palermo.

CONSTITUCION	9618
FLORES	18792
FLORESTA	4596
LA BOCA	5024
LINIERS	7051
MATADEROS	8554
MONTE CASTRO	3035
MONTSERRAT	8404
NUEVA POMPEYA	7645
NUÑEZ	5381
PALERMO	27689
PARQUE AVELLANEDA	5639
PARQUE CHACABUCO	7459
PARQUE CHAS	1336
PARQUE PATRICIOS	6103
PATERNAL	2043
PUERTO MADERO	338
RECOLETA	16608
RETIRO	8488
SAAVEDRA	5459
SAN CRISTOBAL	5989
SAN NICOLAS	17319
SAN TELMO	4986

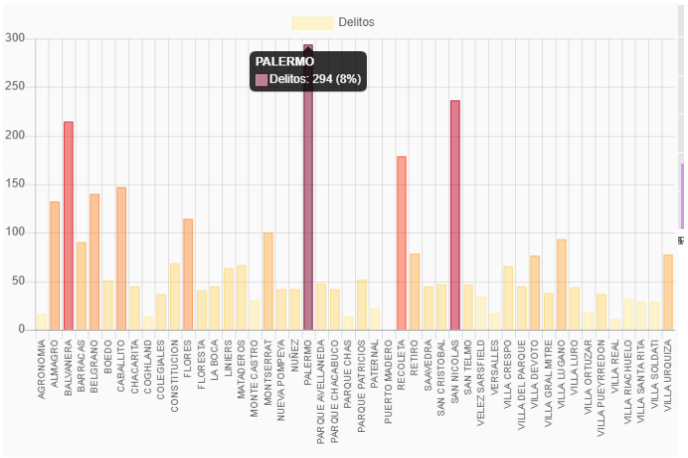


Figura 49 Barrios gráfica de barras.

Fuente: Elaboración propia, 2019

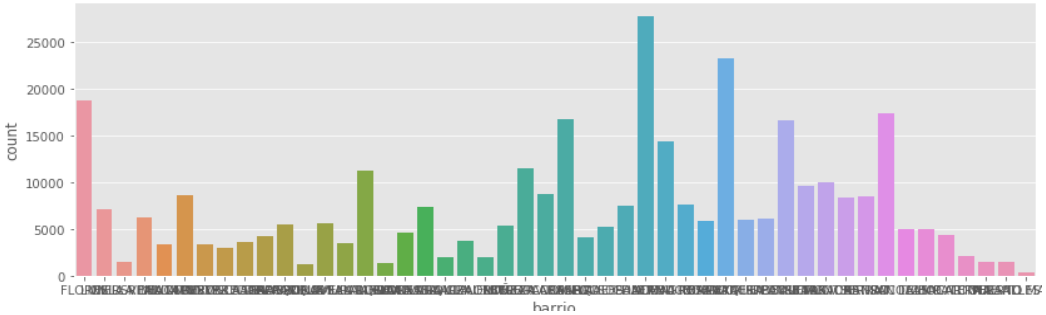


Figura 50 Barrios combinaciones gráfica de barras.

Fuente: Elaboración propia, 2019.

4.1.3. Objetivo N° 3**: Preparar los datos y aplicar algoritmos de clustering a los datos recolectados.**

El agrupamiento con la herramienta Power BI. Clustering por datos de delitos por barrios:

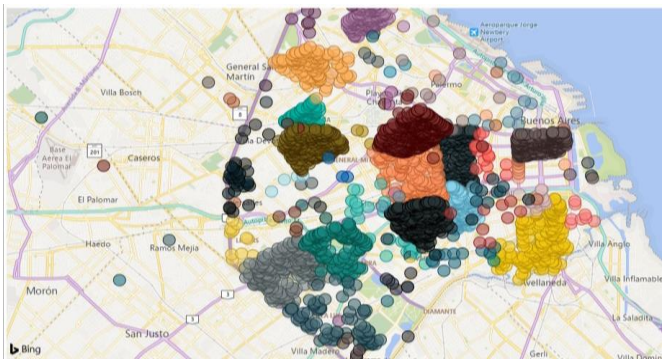


Figura 51 Geo Sectorización de datos

Fuente: Elaboración propia, 2019.

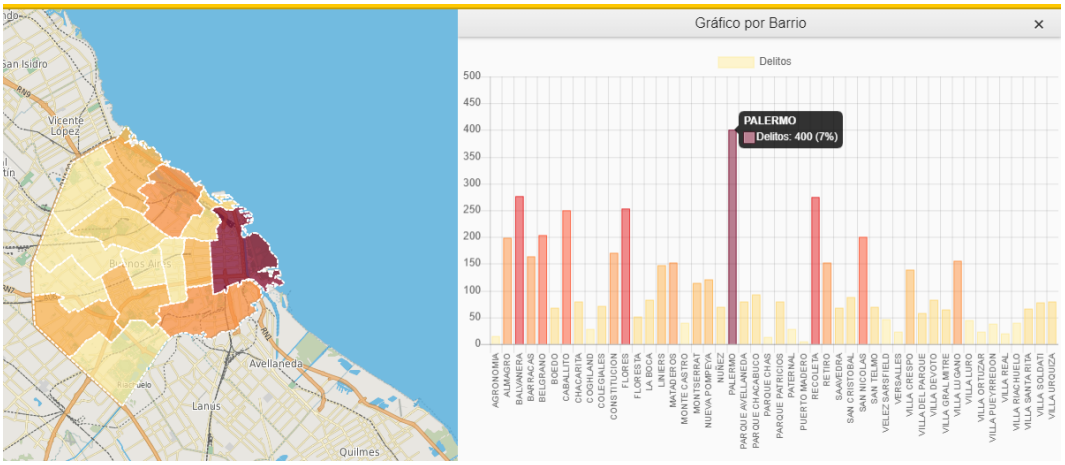


Figura 52 Barra y geo sectorización.

Fuente: Elaboración propia, 2019.

Gráfico de tipo delitos por las comunas y cómo podemos ver la comuna 1 tiene mayor índice de delito.

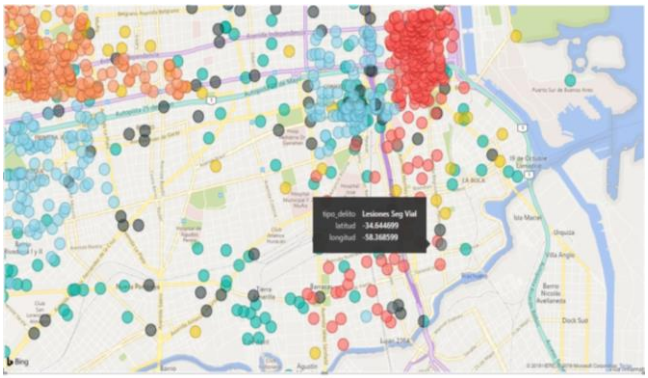


Figura 53 Geo sectorización por puntos en delito.

Fuente: Elaboración propia, 2019.

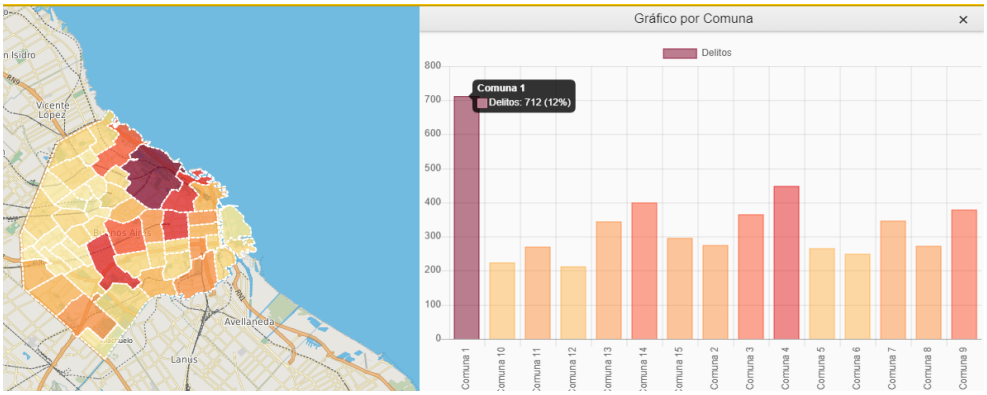


Figura 54 Comuna gráfica de barras y geo sectorización

Fuente: Elaboración propia, 2019.

Gráfico por barrios según los delitos.

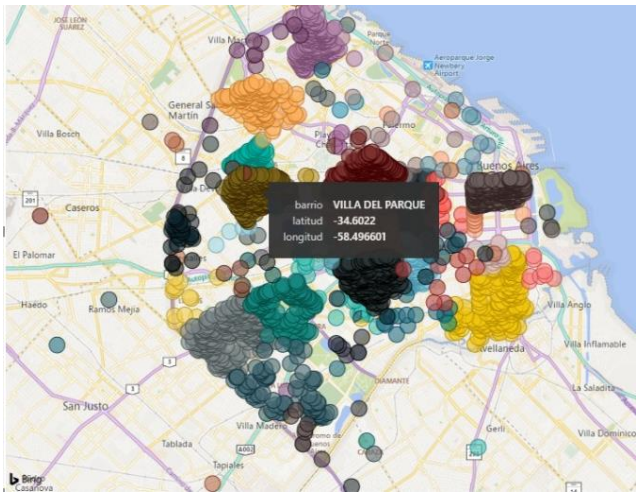


Figura 55 barrios según delitos geo sectorización.

Fuente: Elaboración propia, 2019.

La presente gráfica representa a las horas y días de alto riesgo en donde ocurre los delitos. pudiendo apreciar que los delitos más propicios son los días lunes y marte se las semanas.

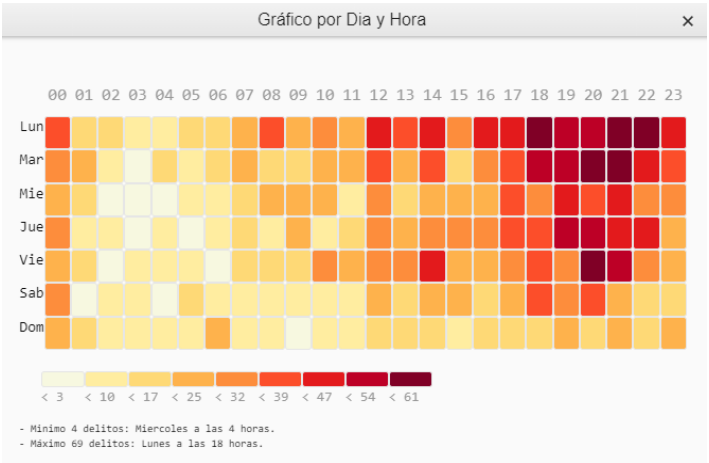


Figura 56 Días y horas.

Fuente: Elaboración propia, 2019.

clustering en Python

Según los análisis previamente realizados obtuvimos que los delitos más mencionados u ocurrido en la ciudad de buenos aires son de: homicidio, hurto y hurto automotor.

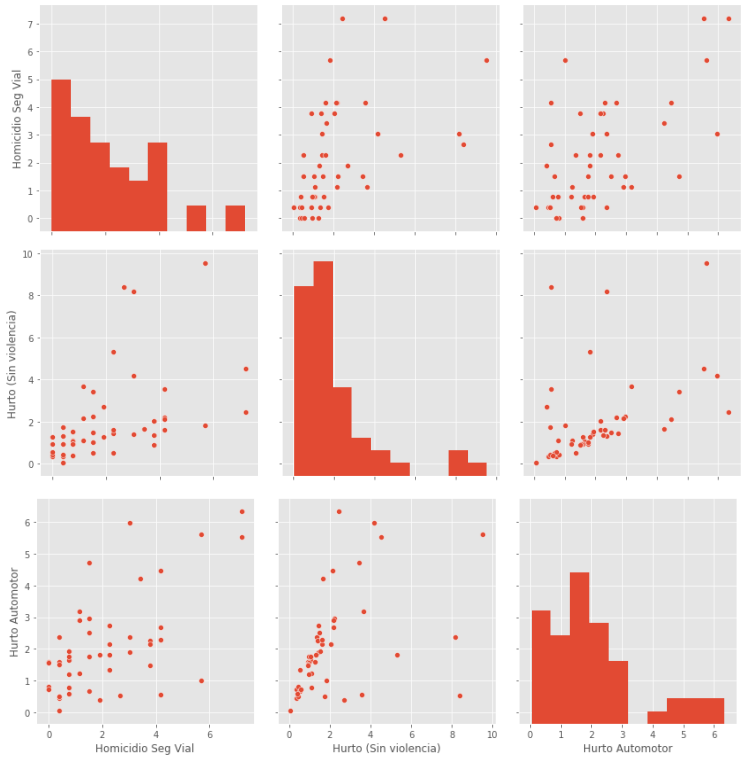


Figura 57 gráfica de barras de tipos de delitos.

Fuente: Elaboración propia, 2019.

Aquí tenemos los datos numéricos del porcentaje de concurrencia, de los delitos analizados en las anteriores figuras.

resumen del agrupamiento de los tipos de delitos con más concurrencia.

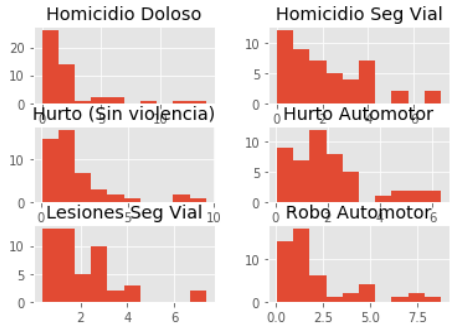


Figura 58 gráfica de barras T_D

Fuente: Elaboración propia, 2019.

ya que los datos a estudio son los delitos concretamos la estructura de datos que utilizaremos para alimentar el algoritmo. según los estudios de análisis de datos antes

realizados podemos ver que los dos tipos de delitos más cometidos en la ciudad de Buenos Aires son los Robo con violencia, hurto sin violencia y homicidio doloso graficamos los clustering que son más cercanos a los centroides.

```
In [70]: X = np.array(dataFrame[["Homicidio Doloso","Hurto (Sin violencia)","Robo (Con violencia)"]])
         y = np.array(dataFrame.index)
         X.shape

Out[70]: (48, 3)

In [90]: fig = plt.figure()
         ax = Axes3D(fig)
         ax.scatter(X[:, 0], X[:, 1], X[:, 2],s=60)

Out[90]: <mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x257aa37cf28>
```

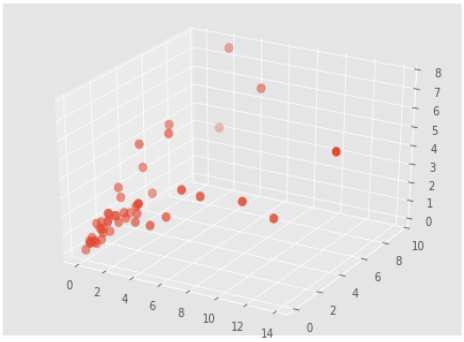


Figura 59 Estructura de datos.

Fuente: Elaboración propia, 2019.

En número de clustering que tomaremos para los estudios lo calculamos mediante la siguiente figura expuesta, para ver el número y el rango. ya que es la selección de número óptimo no supervisado de los grupos.

Obtener el valor K

```
In [91]: Nc = range(1, 20)
         kmeans = [KMeans(n_clusters=i) for i in Nc]
         kmeans
         score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
         score
         plt.plot(Nc,score)
         plt.xlabel('Number of Clusters')
         plt.ylabel('Score')
         plt.title('Elbow Curve')
         plt.show()
```

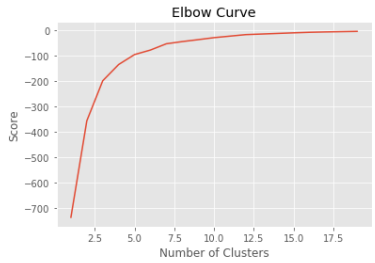


Figura 60 K medias gráfica para centroides

Fuente: Elaboración propia, 2019.

Podemos observar en la siguiente figura que fue calculada 5 centroides para la clusterización de puntos en cuanto al tipo delitos.

```
[[11.35802469  3.22860953  3.84323196]
 [ 0.54673721  1.06460643  1.07050321]
 [ 1.5308642   4.03793429  3.97836641]
 [ 3.62139918  1.58139442  2.30132177]
 [ 3.45679012  8.69432334  5.96416203]]
```

Figura 61 K centroides.

Fuente: Elaboración propia, 2019.

Podemos ver la gráfica 3D con los colores para los grupos y veremos si se diferencia. las estrellas son el centroide de cada grupo de objetos. en cuanto a las dimensiones de colores y sus cantidades son: rojos 3, verde 28, azul 5, turquesa 9, amarillo 3 de las cuales los grupos con más objetos son los verdes que son el **hurto sin violencia**.

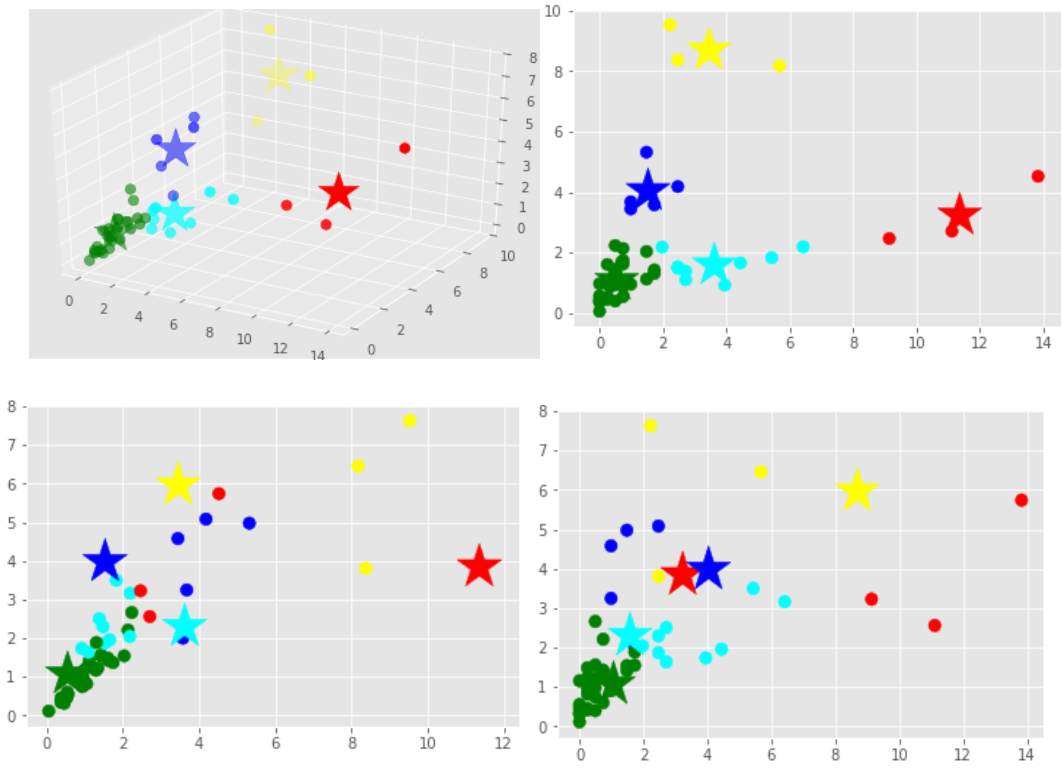


Figura 62 gráficas de clusters según 4 centroides

Fuente: Elaboración propia, 2019.

4.1.4. Objetivo N° 4**: Diseñar la plataforma para la visualización de datos.**

La estructura de la plataforma lo realizamos con el análisis de sistemas PMBOOK.

a. Base de datos.

b. Análisis de requerimientos.

c. Construcción con django y materialize. (aquí agregamos d3 para que pueda integrarse y poder construir, de acuerdo a la alimentación de datos que podamos obtener con el tiempo de las instituciones encargadas de las mismas)

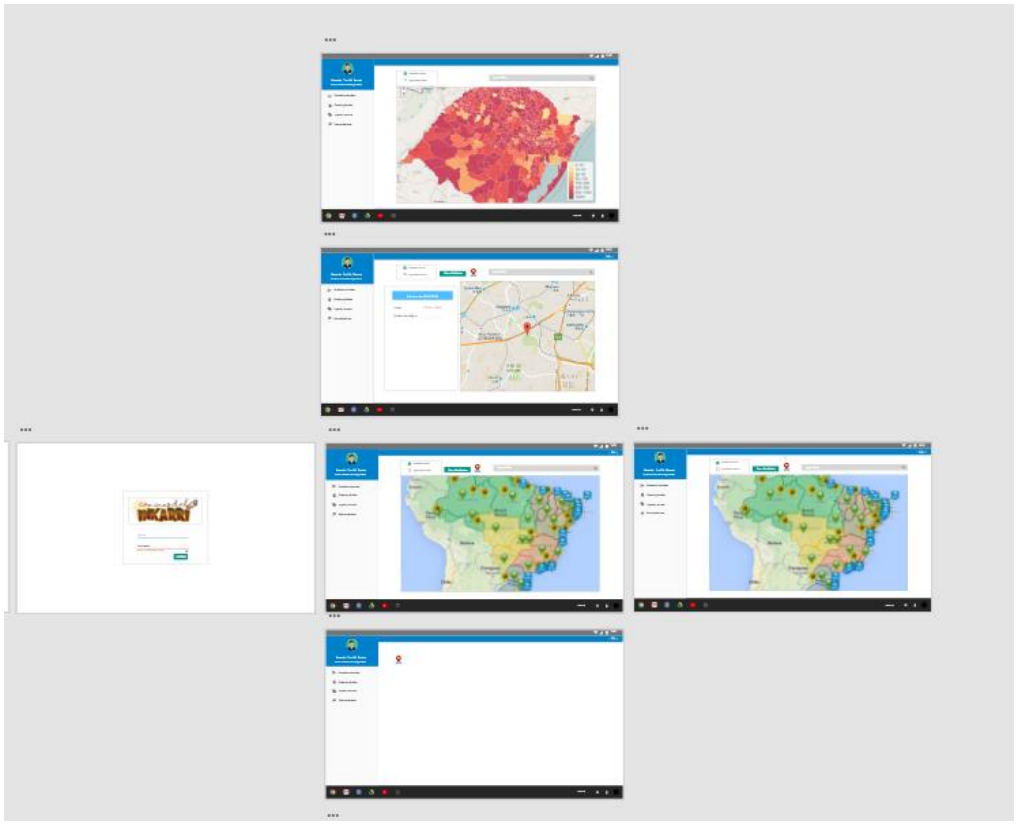


Figura 63 Prototipado del sistema de geo sectorización

Fuente: Elaboración propia, 2019.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En el presente trabajo de investigación podemos concluir que la obtención de datos tanto de la web como los datos estructurados para un buen análisis de datos tienen que ser numéricos y no cualitativos ya que se hace complicada su manipulación de los mismos.

En los análisis de datos se logró diagnosticar que los delitos mayormente cometidos son los de hurto, robo y homicidios en la ciudad de Buenos aires, así como también que el barrio con más índice de actividad delictiva es la de Palermo.

Los datos en análisis son similares a los datos en construcción para el turismo por lo cual concluimos que los estudios realizados serían similares con centrándonos ya no en los barrios si no en los lugares de riesgo delictivo.

5.2. Recomendaciones

- En el presente trabajo de investigación lo más complicado fue obtener datos de la ciudad de puno de la inseguridad turística, se hizo la solicitud de de los datos al centro de PENTUR, pero no obtuvimos respuesta. por lo cual utilizamos datos de una open data en el cual eran datos de la Ciudad de Buenos aires, los datos tomados eran semejantes a las del análisis de la base de datos del PENTUR PUNO es por eso que se tomaron como modelos de estudio. Necesitamos una plataforma en donde liberen datos para poder hacer los análisis y que no sean complicadas los procesos administrativos para su adquisición de los mismos.
- Para el estudio de los datos y su visualización necesitamos datos de las instituciones encargadas, así como también necesitamos que los datos salgan de una aplicación para

que sea fácil su manipulación y alimentación de la plataforma de visualización de datos.

- Se recomienda que los aplicativos de alimentación sean tanto en la web con móviles.
- Recomendamos que todos los datos almacenado sean convertidos en un atributo numérico ya que con eso se hará fácil el análisis o implementación de los datos a algún algoritmo.
- Se recomienda a un futuro implementar en un aplicativo web y móvil la visualización de datos con su respectivo sistema de alimentación para que pueda ser una plataforma de información que sirva tanto al sector público de las instituciones como a los turistas.

REFERENCIAS BIBLIOGRÁFICAS

Al Turista, R. D. P. (2017). Plan de protección al turista. Lima, Perú: Ministerio de Comercio Exterior y Turismo do Perú.

Arancibia, J. A. G. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. Recuperado de http://oldemarrodriguez. com/yahoo_site_admin/assets/docs/Documento_CRISP-DM, 2385037.

Benítez, I. (2005). Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos. España: Universidad Politécnica de Valencia, Departamento de Ingeniería de Sistemas y Automática.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Recuperado de http://www.iidia.com.ar/rgm/CD-TIpEI/TEI-2-CRISP-DM-GdP-material.pdf

Contreras, F. (2016). Introducción a machine learning. Recuperado de https://www.zemsania.com/recursos-zemsania/whitepapers/DTS/Machine_learning.pdf

Díez, R. P., Gómez, A. G., & de Abajo Martínez, N. (2001). Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva. Universidad de Oviedo.

Galipienso, M. I. A., Quevedo, M. A. C., Pardo, O. C., Ruiz, F. E., & Ortega, M. A. L. (2003). Inteligencia artificial: modelos, técnicas y áreas de aplicación. Editorial Paraninfo.

Garmendia, I. (2017). Plan de protección al turista.

Gloria, G. m. (2017). Análisis del impacto económico anual del WTTC. Recuperado de: https://www.wttc.org/research/economic-research/economic-impact-analysis/

Gloria, G. m. (2017). Impacto del sector de viajes y turismo en las ciudades 2017 américa latina. Recuperado de: https://www.wttc.org/-/media/files/reports/economic-impact-research/cities-2017---regional/latin-america-city-travel-and-tourism-impact-2017-spanish.pdf?la=en

Jaulis Rua, J. J., & Vilcarromero Giraldo, J. R. (2015). Sistema de predicción de hechos delictivos para la mejora del proceso de prevención del delito en el distrito de La Molina utilizando minería de datos.

Juan Carlos (3 de enero del 2018), Portal de turismo. Seguridad, estrategia y oportunidad: 3 retos del sector turismo para el 2018 Recuperado de

http://www.portaldeturismo.pe/noticia/seguridad-estrategia-y-oportunidad-3-retos-del-sector-turismo-para-el-2018-editorial-

Montes Ipenza, D., Rodriguez, A. J., Zamora Aguilar, R., & Zambrano Chávez, E. (2017). Investigación para la implementación de una empresa de servicios turísticos asociados a la pesca deportiva en el Perú.

Rodríguez León, Ciro, & García Lorenzo, María Matilde. (2016). ADECUACIÓN A METODOLOGÍA DE MINERÍA DE DATOS PARA APLICAR A PROBLEMAS NO SUPERVISADOS TIPO ATRIBUTO-VALOR. Revista Universidad y Sociedad, 8(4), 43-53. Recuperado en 02 de octubre de 2018, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005&lng=es&tlng=es.

Rodríguez León, Ciro, & García Lorenzo, María Matilde. (2016). ADECUACIÓN A METODOLOGÍA DE MINERÍA DE DATOS PARA APLICAR A PROBLEMAS NO SUPERVISADOS TIPO ATRIBUTO-VALOR. Revista Universidad y Sociedad, 8(4), 43-53. Recuperado en 20 de septiembre de 2018, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005&lng=es&tlng=es.

Ríos, S., Hinojosa, C., & Delgado, R. (2013). Aplicación de la metodología openup en el desarrollo del sistema de difusión de gestión del conocimiento de la espe, 10.

Sánchez, I. J. B. Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos.

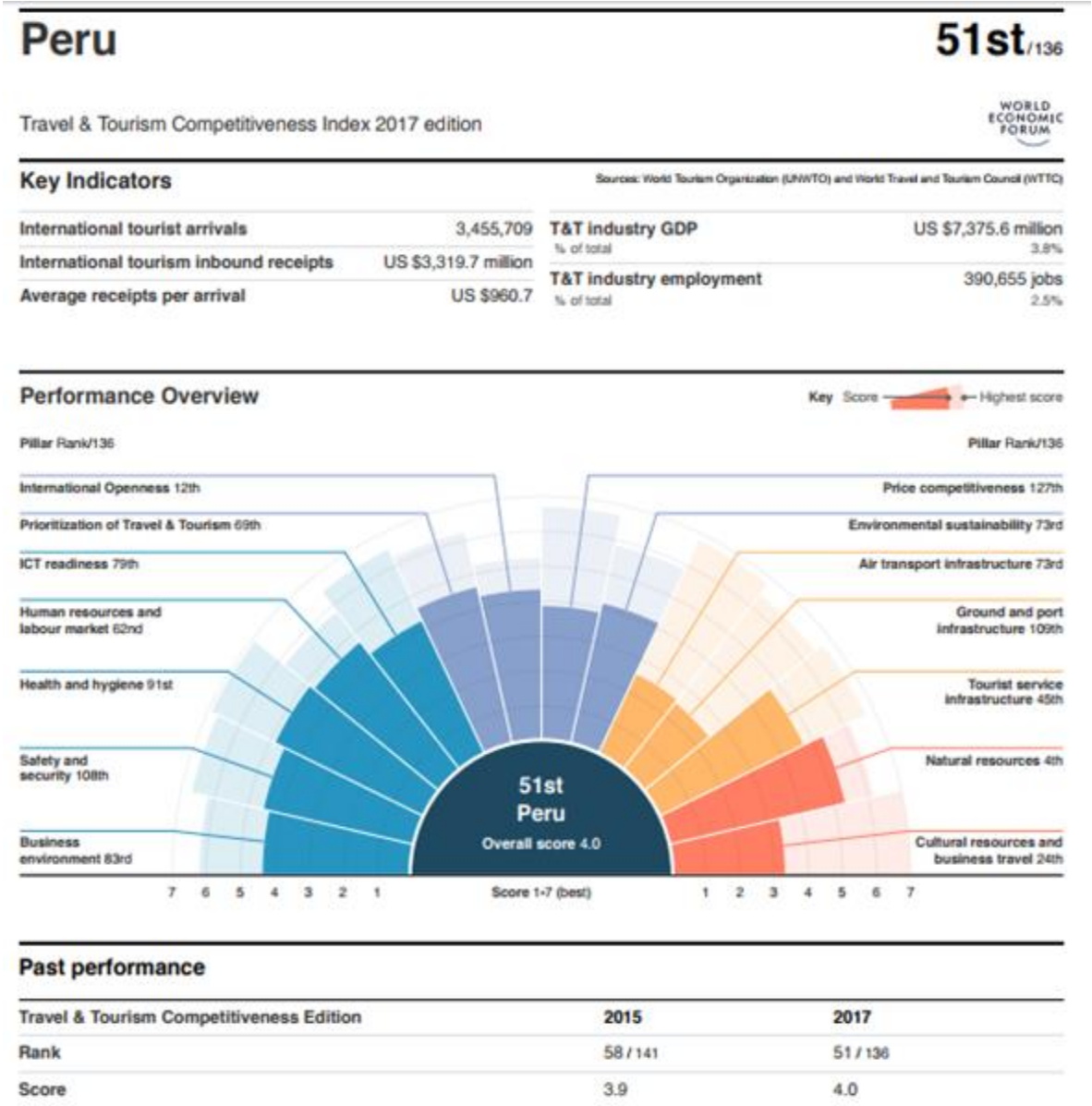
Riquelme, J.C., Ruiz, R. & Gilbert, K. (2006). Minería de datos: conceptos y tendencias, inteligencia artificial. Revista Iberoamericana de Inteligencia Artificial, 29 pp. 11-18.

López, J. (2018). Web scraping.

ANEXOS

Anexo 1: **Estadística** de la WEF

el informe completo del al WEF TTCR 2017 puede ser visualizado en: http://www.cdi.org.pe/InformeGlobaldeViajesyTurismo/doc/2017/WEF_TTCR_2017_web_0401.pdf



Peru

51st/136

Travel & Tourism Competitiveness Index 2017 edition

WORLD ECONOMIC FORUM

Key Indicators

Sources: World Tourism Organization (UNWTO) and World Travel and Tourism Council (WTTC)

International tourist arrivals	3,455,709	T&T industry GDP	US $7,375.6 million
International tourism inbound receipts	US $3,319.7 million	% of total	3.8%
Average receipts per arrival	US $960.7	T&T industry employment	390,655 jobs
		% of total	2.5%

Performance Overview

Past performance

Travel & Tourism Competitiveness Edition	2015	2017
Rank	58 / 141	51 / 136
Score	3.9	4.0

Anexo 1 Estadística de la WEF

Anexo 2: **Mapic del proyecto de investigación**

VARIABLE FÁCTICA	DIMENSIONES	INDICADORES
1. Inseguridad ciudadana en el contexto turístico.	.1. Seguridad	1.1.1 Seguridad ciudadana 1.1.2 Seguridad Turística instituciones y funciones, citas
VARIABLE TEMÁTICA	**EJES TEMÁTICOS**	**SUB EJES TEMÁTICOS**
2. Algoritmos de clustering 3. Web scraping	2.1 Algoritmos de clustering 2.2 Minería de datos 3.1 Web scraping y minería de datos	2.1.1 IA y NLP 2.1.2 Técnica de agrupamiento y reconocimiento de patrones turismo patrones 2.1.3 Técnicas de agrupamiento para datos cuantitativos y cualitativos 2.2.1 Big data 2.2.2 Base de datos estructurados y no estructurados para minería de datos. 2.2.3 Algoritmos de minería de datos. 3.1.1 Maquetación web 3.1.2 Extracción de datos 3.1.3 tipos
VARIABLE PROPOSITIVA	**EJES PROPOSITIVOS**	**SUB EJES PROPOSITIVOS**
4. Prototipo de Sistema de geo – sectorización	4.1 Análisis y diseño 4.2 Prototipo 4.3 Implementación de mapas Geo localizadas (sectorización)	4.1.1 Análisis y toma de requerimientos 4.3.1 Desarrollo del Prototipo 4.3.2 Utilización de Arcgis.

Anexo 2 Mapic del proyecto de investigación