INFOTEC POSGRADOS	Fecha: 07/04/2022
Solicitud de Registro de Tema para Titulación	

Quien suscribe, estudiante regular de la Maestría en Ciencia de Datos e Información

Nombre:	**FRANCISCO EDUARDO ZAVALA RODRIGUEZ**

Solicitamos ante el Coordinador Académico de la Maestría en Ciencia de Datos e Información, la autorización y registro para titulación en la Maestría, en la modalidad:

	Reporte Analítico de Experiencia Laboral
	Propuesta de Intervención
x	Implementación de un Proyecto
	Solución Estratégica

TITULO TENTATIVO DEL PROYECTO	Reconocimiento de texto para clasificación de causas básicas CIE-10

INFOTEC POSGRADOS	Fecha: 07/04/2022
Solicitud de Registro de Tema para Titulación	

Introducción.

El ISSSTE es una dependencia gubernamental que brinda no únicamente servicios para pensiones para trabajadores de gobierno si no que también una de sus principales funciones es brindar un servicio de calidad a los derechohabientes que acuden a sus unidades médicas las cuales están divididas en tres grandes grupos:

1. Unidades de primer nivel, dentro de las cuales se encuentran UMF (unidades de medicina familiar),CMCT (centros médicos en centros de trabajo),CMF(clínicas de medicina familiar) y CAF (centros de atención familiar),por mencionar algunas dentro de esta unidades, una de sus principales funciones es detener y prevenir a futuro enfermedades crónicas degenerativas así como también atender padecimientos comunes como lo son gripas, problemas de nutrición, planificación familiar, caries, etc.
2. Unidades de segundo nivel, como son HG (hospitales generales), CH (clínicas hospital), CMFEQ (clínicas de medicina familiar y quirófano) y CE (clínicas de especialidad) en estas unidades se atienden a pacientes que no pudieron ser controlados en las unidades de primer nivel o que requieren alguna intervención quirúrgica simple, fracturas, urgencias menores o alguna atención especial de algún padecimiento presentado por el derechohabiente.
3. Unidades de tercer nivel aquí se encuentran HR Y CMN (centro médico nacional 20 de noviembre), estas unidades únicamente atienden padecimientos crónicos degenerativos o pacientes con problemas grabes de salud, como puede cáncer, diabetes mellitus en fase terminal, problemas cardiacos, trasplantes, casos no reconocidos o de estudio de alguna enfermedad.

Todas estas unidades generan información diaria sobre padecimientos, muertes, estudios, consultas médicas, accidentes, personal que labora, etc. y si esta información no tiene un buen tratamiento puede ocasionar problemas como desabasto de medicamento, la mala adquisición de equipo médico, mala contratación de personal médico (enfermeras, médicos especialistas paramédicos, personal administrativo, etc.).

Por ende, la información y un buen análisis de estos datos que se generan ayudan a la institución a poder hacer cada día mejor su trabajo.

Objetivo

Como principal objetivo de este problema se busca crear un algoritmo que se suficientemente eficaz para reconocimiento de texto e imagen que nos pueda clasificar padecimientos originados del diagnóstico médico capturado en un certificado de defunción y otorgar una mejor clasificación CIE-10(Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud)

y compararla con el sistema que actualmente usado por el instituto llamado “IRIS”(Sistema de Codificación Automatizada de causa de muerte)

Antecedentes

La problemática surge de la revisión y validación de médica enfocada a muertes maternas, fetales, defunciones(padecimiento por el cual una persona muere) y morbilidades (padecimiento o causa de ingreso de un paciente a un hospital o unidad médica), sin embargo el mal diagnostico elaborado por un sistema llamado “IRIS” ya que este se basa en el diagnóstico del médico capturado en el certificado de defunción y según las causas básicas asignadas en él, sin embargo al hacer la revisión de estos datos, resulta que el diagnóstico realizado por el médico y la causa básica de muerte no coincide con la dictada por el programa ya antes mencionado.

Esto ocasiona que la confiabilidad del sistema a realizar codificaciones por medio de la CIE-10 no tenga mucha valides y se tenga que revisar dato por dato comparando el diagnóstico del médico contra el del sistema, ya que, si esto no se revisa adecuadamente, al presentar la información sin un proceso previo de validación, tanto las estadísticas que se alimentan de estos datos, así como indicadores serían erróneos.

Los modelos supervisados tienen la característica de utilizar datos previamente clasificados estos datos a su vez se entrenan para realizar tareas específicas y se utilizan para predecir resultados ya conocidos, un ejemplo claro, podría ser la clasificación de animales de 2 y cuatro patas , para esto tendríamos que tener datos previamente clasificados, donde indique si el animal es de 2 o 4 patas, una de las ventajas de estos modelos es que “aprenden ” de datos históricos de entrenamiento y se ocupen en datos de desconocidos para obtener la salida correcta.

Algunos de los modelos más concurridos dentro del aprendizaje supervisado son:

- Árboles de decisión
- Clasificación de Naïve Bayes
- Regresión por mínimos cuadrados
- Regresión Logística

Árboles de decisión

Los árboles de decisión son herramientas de clasificación, que permite realizar secuencias basadas en el uso de resultados y probabilidades asociadas, también son utilizados para generar sistemas expertos, búsquedas, arboles de juegos etc. Solo por mencionar algunas de sus basta aplicaciones.

Algunos de las propiedades de los árboles de decisión son:

- Reduce el número de variables
- Permite la clasificación de nuevos casos siempre y cuando no existan modificaciones sustanciales en las condiciones bajo las cuales se realizaron las iteraciones
- Los árboles de decisión no son lineales lo que significa que hay flexibilidad para poder explorar, planificar y predecir varios resultados posibles para cada curso de acción
- Son objetivos ya que estos se centran en la probabilidad y los datos, no en las emociones.

Clasificación Naive Bayes

Estos algoritmos de aprendizaje supervisado y de machine learning, se basan en una técnica de clasificación estadística “teorema de Bayes”, dentro de estos modelos se asume que las variables predictoras son independientes entre sí, ya que proporcionan una manera más fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad.

Esto se consigue gracias a que se proporcionan probabilidades (la probabilidad posteriori de que cierto evento A ocurra dadas algunas probabilidades de eventos anteriores “apriori”)

Regresión por mínimos cuadrados

El método de mínimos cuadrados se aplica para ajustar rectas a una serie de datos presentados como el punto en el plano, de esta manera los mínimos cuadrados nos proporcionan un criterio con el cual podremos obtener la mejor recta que representa a los puntos dados y de está manera ajustar la recta que más se ajuste a los datos presentados y poder realizar proyecciones futuras sobre los mismos datos trabajados.

Regresión logística

La regresión logística es una técnica multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica (valores ceros y unos, si y no, etc.) y un conjunto de variables independientes métricas y no métricas.

Uno de los objetivos principales de la regresión es de modelar como influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular, de este mismo se derivan dos objetivos sistemáticos.

- Investigar cómo influye en la probabilidad de ocurrencia de un suceso, la presencia o no de diversos factores y el valor o nivel de estos
- Determinar el modelo más parsimonioso y mejor ajustado que describa la relación entre las variables respuesta y un conjunto de variables regresoras.

Ya una vez teniendo una vaga idea sobre los modelos de aprendizaje supervisado y algunos de sus modelos más representativos, vamos ahora hablar de los modelos no supervisados, los cuales a diferencia de los modelos supervisados estos no cuentan con datos previamente clasificados, lo que los hace un poco más complejo dentro estos modelos se pretende experimentar y tratar de interpretar el resultado del modelos según el diseño dado.

En resumen, el objetivo principal es estudiar la estructura intrínseca de los datos, estas técnicas se pueden condensar en dos tipos principales de problemas que el aprendizaje no supervisado trata de resolver, como los son agrupación y reducción de dimensionalidad.

Algunos de estos modelos son:

- Algoritmos de clústering
- Análisis de componentes principales.

Algoritmos de clústering

Estos algoritmos muy conocidos como clústering o algoritmos de clasificación, su principal función y como su nombre lo dice es encontrar diferentes grupos dentro de los elementos de los datos. Para ello, los algoritmos de agrupamiento encuentra la estructura en los datos de manera que los elementos del mismo grupo sean más similares entre sí que con los de clúster diferentes, algunos de los modelos más utilizados son:

- K-means
- Clúster jerárquico
- Density based scan clustering
- Modelo de agrupación Gaussiano

Análisis de componentes principales

Este modelo es un modelos estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información, este a su vez tiene como objetivo de predecir una variable de respuesta dada una serie de predictores,

para ellos se disponen de n características y de la variable de respuesta, este modelos permite reducir la información aportando miles de variables en solo unas pocas componentes lo que hace a este modelo como uno muy útil de aplicar otras técnicas estadísticas como pueden ser clúster, regresiones, etc.

2.2. **Planteamiento del problema**.

Para poder solucionar este problema lo que se plantea es elaborar un programa o un algoritmo capaz de reconocer texto y palabras clave que nos ayuden a asignar un diagnóstico correcto con base en la causa básica de muerte capturada en el certificado de defunción por el médico y así asignar una correcta clasificación CIE-10 posteriormente medir la eficacia del algoritmo y compararlo con lo dictado por el sistema IRIS

2.3. **Objetivo General**

El objetivo general de este trabajo de tesis es encontrar el mejor clasificador para la búsqueda y coincidencia de los padecimientos presentados por los derechohabientes dentro de las unidades de segundo y tercer nivel, donde los principales errores se cometen al hacer la captura de la información registrada dentro de los certificados de defunción y dentro del sistema interno del instituto.

Para esto nos apoyaremos en el sistema internacional IRIS desarrollado en Alemania en 2007, el cual realiza las clasificaciones de morbilidades y moralidad por medio de catálogo registrados en su sistema y por medio de algoritmos de inteligencia artificial para realizar el mejor diagnóstico.

Objetivos específicos.

Además, con ello poder observar y comparar los resultados obtenidos con el algoritmo elaborado de manera interna y poder tomar una muestra para comparar los resultados obtenidos de la realización de las clasificaciones de mortalidad y morbilidad y poder definir cuál de los dos algoritmos es mejor que el otro

3. Resultados Esperados

Como uno de los principales resultados a los que se pretende llegar es que el clasificador elaborado tenga mejor tasa de aciertos sobre la clasificación de enfermedades y muertes que se registran dentro del instituto.

4. Programa de Trabajo

Este proyecto pretende llevarse en 5 etapas o fases

1. Fase de investigación
2. Fase de recolección de datos
3. Fase de análisis de datos
4. Fase de presentación de dato
5. Fase de resultados

Metodología

La maquinas de soporte vectorial mejor conocidas como ***Support vector machine*** son un conjunto de algoritmos de aprendizaje estadístico supervisado pertenecientes a la familia de los clasificadores lineales desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T en el año 1995.

Sin pérdida de generalidad, suponiendo que tenemos ejemplos de sólo dos categorías una SVM construye un hiperplano en un espacio de dimensionalidad muy alta o incluso infinita. Este hiperplano separa de forma óptima los puntos de una clase de la de otra. En el concepto de "separación óptima" es donde reside la característica fundamental de las SVM, se busca el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo.

Las SVM se desarrollaron inicialmente para resolver problemas de clasificación. Sin embargo, se han reformulado de múltiples formas a lo largo de los años, sus variantes más conocidas son: SVM para regresión, SVM para resolución de ecuaciones integrales, SVM para estimar el soporte de una densidad, SVMs que usan diferentes costes de margen blando y parámetros.

También se han ensayado otras formulaciones del problema dual. Las técnicas de optimización cuadrática son fundamentales a la hora de resolver estos problemas. Actualmente, las SVM tienen utilidad en una extensa variedad de áreas. Algunas de sus aplicaciones más importantes son el reconocimiento de caracteres, detección de intrusos, reconocimiento del habla o la bioinformática.

Como acabamos de ver este modelo tiene diversas aplicaciones por lo cual nos será de mucha utilidad para elaborar y clasificar los egresos y defunciones por medio de las codificaciones de la cie-10.

Para poder proceder a la realización del modelo de clasificación lo primero que se procederá a realizar es un catálogo de clasificaciones de la cie-10 [https://ais.paho.org/classifications/chapters/pdf/volume1.pdf] está lista contiene todos los padecimientos a nivel mundial, a los cuales se les asigna una clave según el tipo de padecimiento por mencionar algún ejemplo veremos cómo están clasificados los tumores malignos, dentro de la CIE-10 se encuentran codificados con la letra “C00-C097”.

Como primera instancia ya una vez realizado el catálogo según la normatividad lo que se pretende realizar es un conjunto datos ya clasificados según su causa básica de muerte estas clasificaciones son elaboradas por expertos codificadores que ayudan a facilitar el trabajo del sistema IRIS.

Otro modelo para utilizar será el modelo de clústering k-menas el cual nos ayudara a validar si las clasificaciones de la cie-10 de los certificados de defunción están correctamente clasificados y se les asignara un grupo según sea el padecimiento.

Primero seleccionamos los k centroides iniciales $\{C_1, \dots, C_K\}$, ya una vez teniendo los centroides de cada dato procedemos a asignar los objetos x_i del conjunto de datos X a su centroide más cercano en este caso a su clasificación correspondiente según la CIE-10 ya procesada en un catálogo, recalculamos los nuevos centros, regresamos al paso donde asignamos cada elemento a su centro hasta que el algoritmo converja.

Para poder validar que la información fue clasificada correctamente utilizaremos dos índices de validación.

Inidice S_Dbw

Este índice evalúa los resultados de un algoritmo de agrupamiento, en función de la densidad y separación de los grupos. Para lo cual mide la varianza intra-grupo e inter-grupo.

$$S_{dbw} = Scatt + Dens\_bw$$

Indice Ps

Este índice calcula el promedio de la distancia simétrica hacia otros centros.

$$PS(C) = \frac{1}{K}\sum\left[\frac{1}{P\_i}\sum_{j=i}^{p\_i}\frac{d\_c(x_i, c_i)}{d\_min}\right]$$

Donde $d_c(x_i, c_i) = d_s(x_i, c_i) d_e(x_i, c_i); d_e = (x_i, c_i)$ es la distancia euclidiana entre el punto $x_i \; y \; el \; c_i \; y \; d_s$ es la distancia simétrica.

Una vez hechas las validaciones con ayuda del Support vector machine en una de sus aplicaciones de análisis de texto comenzaremos a realizar la parte de clasificación de palabras a las cuales les asignaremos una clasificación ya validada por el k-means, para seleccionar el mejor modelo que compararemos contra el sistema de codificación IRIS, realizaremos una clasificación con el modelo SMV.

Este modelo tiene dos vertientes los métodos indirectos y los métodos directos.

Métodos indirectos

Los problemas de clasificación múltiple se suelen descomponer en una sucesión de problemas binarios de forma que podemos aplicar el método estándar de resolución de SVM a cada uno por separado. Los dos métodos más representativos de esta técnica son uno frente al resto (1VR) y uno frente a uno (1V1). Ambos métodos son casos particulares de los Códigos Correctores de Errores de Salida

Métodos directos.

Los métodos directos de clasificación múltiple basados en SVM enfocan el problema a un único proceso de optimización, a través de combinar problemas de clasificación binaria en una única función objetivo, de forma que se logra simultáneamente una clasificación de todas las clases. Sin embargo esto conlleva una mayor complejidad a nivel computacional debido al tamaño del problema de optimización cuadrática resultante.

En este caso utilizaremos el método directo por medio del método Weston y Watkin

El cual propone que para métodos de clasificación de k-clases se diseñe una función objetivo que entrene a las k vectores binarios simultaneamentr y maximice los márgenes de cada clase con el resto de ejemplos.

Min

$$\min \frac{1}{2} \sum \|w_m, \|^2 + C \sum \sum \delta_{i,j}$$

$$s.a \; w_{yi} \; \emptyset(x_i)^t + b_{yi} \leq \; {w_{yi}}^t \emptyset(x_i) + b_i + 2 - \delta_{i,j} \; , \delta_{i,j} > 0 \; , i = 0, \dots, l \; , t \in 1, \dots, k \; / \; y_i$$

Visualización.

Para esta parte del proyecto de investigación se analizaron tres gráficos que pueden ayudar a entender mejor el problema de clasificación para esto empezaremos con unas breves descripciones y ejemplificaremos con imágenes la utilidad de las gráficas seleccionadas para el proyecto.

Gráfico de dispersión

Los diagramas de dispersión son una forma eficiente de visualizar información sobre:

- Tendencia
- Concentración
- Clasificaciones
- Valores atípicos dentro de los datos

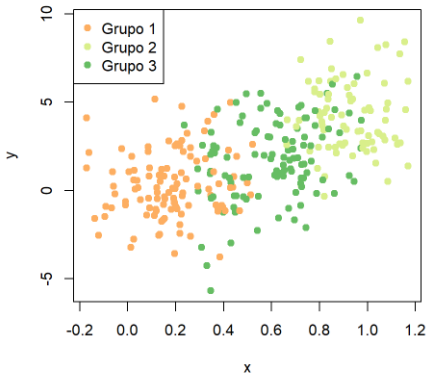
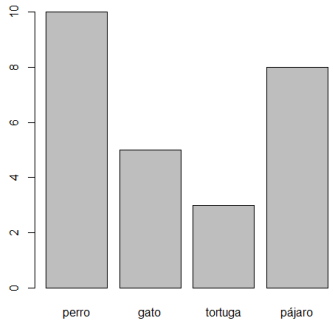


Gráfico de barras

Este tipo de graficas por muy sencillas que parezcan son de mucha utilidad ya que en estas podemos mostrar datos numéricos que se dividen ordenadamente en distintas categorías para que pueda ver tendencias rápidamente en sus datos.



Histograma

Use histogramas cuando deseamos ver cómo se distribuyen sus datos en los grupos; Al agrupar sus datos en estas categorías y luego trazarlos con barras verticales en un eje, verá la distribución de sus calabazas de acuerdo con el peso, también se puede usar histogramas para probar distintos enfoques a fin de asegurarse de crear grupos que se equilibran en peso y son relevantes para su análisis.

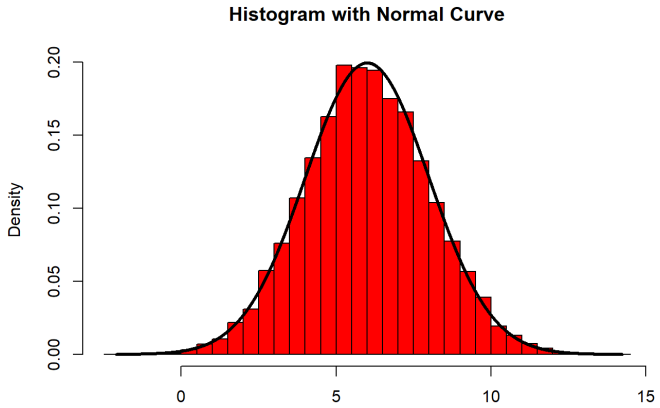


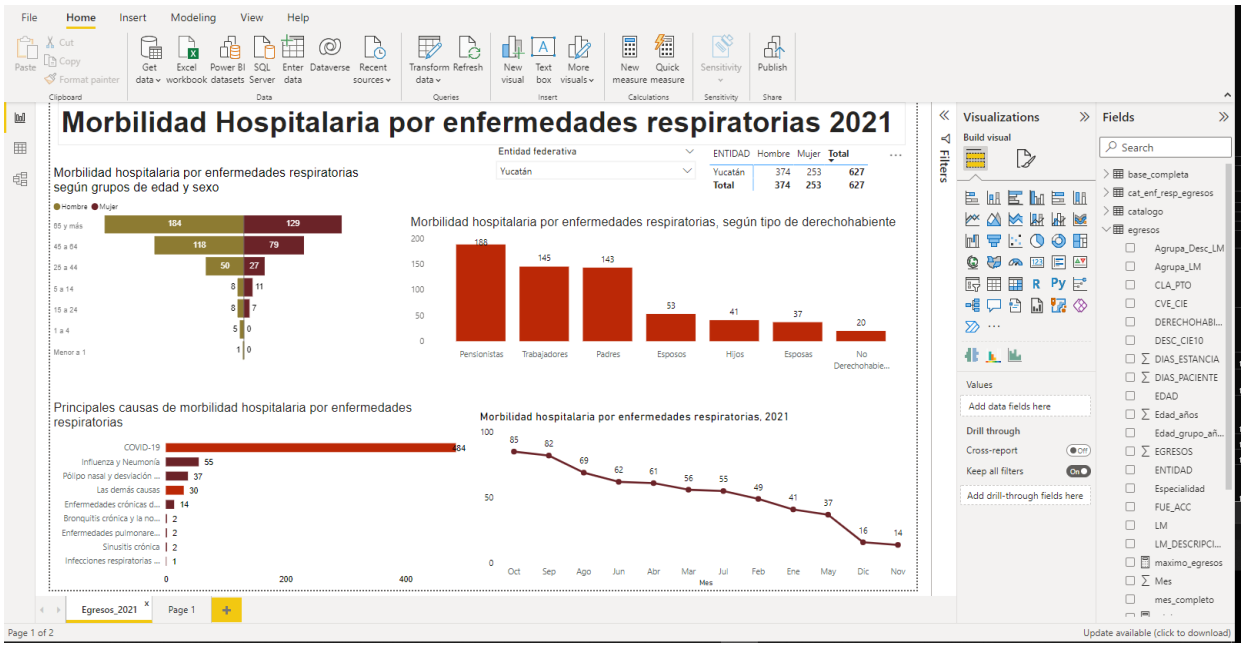
Figura 3: Histograma con la función de densidad de la distribucion normal

Como hemos visto estas no son las únicas gráficas existen más, pero para el proyecto solo utilizaremos estas para mostrar la clasificación de la información visualizar la normalidad de los datos y conocer su distribución estadística para saber de que manera operaremos los datos.

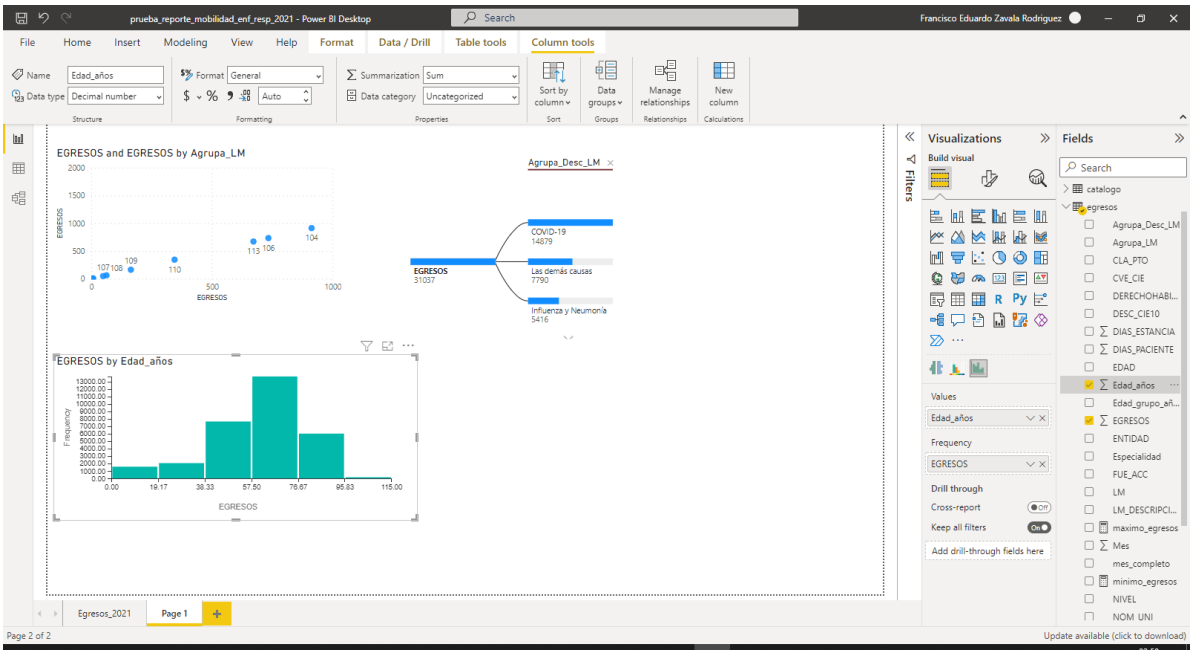
Para la elaboración de estas graficas utilizaremos paqueterías propias de Python, las cuales nos ayudaran para toda la parte de la visualización

- Matplotlib aqueteria especialidad para hacer gráficos de alta calidad
- scatter() comando que nos ayudara para hacer gráficos de dispersion
- hist() comando que nos ayudará para hacer gráficos de histograma
- bar() comando que nos ayudará par ahacer gráficos de barra.

Para la visualización de los datos también haremos uso de la siguiente herramienta llamada Power BI, una herramienta interactiva y muy intuitiva que no ayudara visualizar y encontrar errores en los datos, adjunto un ejemplo del dashboard.



Este nos ayudara para la revisión y análisis de la información, sin embargo para la parte de la clasificación usaremos este otro tipo de dashboard.



Bibliografía consultada para la elaboración de la metodología

- https://www.leanconstructionmexico.com.mx/post/%C3%A1rbol-de-decisiones-ejemplos-de-ventajas-y-pasos-a-seguir
- https://retos-directivos.eae.es/arbol-de-decisiones-ejemplos-de-ventajas-y-pasos-a-seguir/
- https://www.utm.mx/~jahdezp/archivos%20estructuras/DESICION.pdf
- https://ccc.inaoep.mx/~esucar/Clases-mgp/Proyectos/MGP_RepProy_Abr_29.pdf
- https://zaguan.unizar.es/record/59156/files/TAZ-TFG-2016-2057.pdf
- https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f
- https://sistemas.fciencias.unam.mx/~erhc/calculo3_20171/derivadas_parciales_direccionales_2016_12.pdf
- https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc
- https://sisbib.unmsm.edu.pe/bibvirtualdata/Tesis/Basic/Salcedo_pc/enPDF/Cap2.PDF
- https://www.cienciadedatos.net/documentos/35_principal_component_analysis
- https://ais.paho.org/classifications/chapters/pdf/volume1.pdf
- https://rcs.cic.ipn.mx/2016_128/RENTOL_%20Un%20algoritmo%20de%20agrupamiento%20basado%20en%20K-means.pdf
- https://ccc.inaoep.mx/~esucar/Clases-mgp/Proyectos/MGP_RepProy_Abr_29.pdf
- https://powerbi.microsoft.com/en-us/