Primero que nada, aclaro que la base de datos seleccionada por mí corresponde a la distribución geográfica (Especificando la localidad censada), por grupo etario (agrupando en grupos de 5 años, hasta los 85 años y tomando todos los mayores de 85 años como un solo grupo) y estado de salud en Irlanda para el año 2022.

Aunque destaco que en mi análisis realicé algunas modificaciones a la misma a fines de hacerla más funcional, por ejemplo, al utilizar el proceso de Exploratory Data Analysis pude notar que la base estaba incompleta ya que parte de la población se encontraba fuera del censo, y otro caso particular que vale la pena destacar es que en los casos para pueblos de menos de 1500 habitantes no contamos con la ubicación geográfica por condado, por lo que los mantuve fuera de mi análisis ya que mi ides es la de realizar un estudio que permita ayudar a determinar en qué lugar (condado) resultaría mas efectivo destinar por ejemplo inversiones en temas de salud pública.

Para el tratamiento de mi base de datos en Jupyter, el primer paso realizado correspondió a la carga de las librerías que considero necesarias para el manejo, tratamiento estadístico y posterior visualización grafica de la información, por su puesto que a lo largo de todo el proceso fui agregando librerías a medida que fueron siendo necesarias:

1. import pandas as pd: la cual permitirá leer, manipular y analizar los datos presentes en la tabla (.CSV).
2. import seaborn as sns: esta me permitirá mostrar los datos de una manera grafica.
3. from scipy import stats: me permitira realizar el análisis estadístico necesario por ejemplo para el Z-TEST, en mi caso lo utilice para determinar si la diferencia entre hombres y mujeres en el estudio podía representar una variación significativa como para ser tomada en cuenta.
4. import numpy as np: esta me permitira tratar los datos como Array en el caso que sea necesario, dependiendo del análisis que quiera realizar
5. import matplotlib.pyplot as plt: esta biblioteca es la que me permite graficar los datos obtenidos después de modelar la información.
6. from sklearn.naive\_bayes import GaussianNB : sklearn es una biblioteca de machine learning que me facilitara herramientas para diferentes propósitos. En este caso este modulo me permitira aplicar un método de machine learnig para predecir datos en el modelo que estoy generando.
7. from sklearn.model\_selection import train\_test\_split: este módulo me permite que en los algoritmos de machine learning el sistema divida la información en subconjuntos y realice pruebas aleatorias y aprenda para realizar predicciones más precisas cada vez.
8. from sklearn import metrics: este me permite evaluar la precisión de los sistemas de machine learning así como otras métricas en caso de que sean necesarias para la evaluación del mismo.
9. from sklearn.tree import DecisionTreeClassifier: Este módulo me ayudara a aplicar el árbol de decisiones para mi modelo como segundo método de machine learning.

Posteriormente a la carga de las librerías, mi siguiente paso fue cargar la base de datos, e imprimir la misma a fin de conocer su forma y principales características para determinar los pasos a seguir.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Luego ya con la base cargada, note que en la misma se podía encontrar el total de habitantes por cada ciudad, por cada rango etario y por cada estado de salud, pero además una totalidad por cada uno de los campos antes mencionados, así que mi siguiente paso fue filtrar esta información para conocer en un primer momento la totalidad, trabajar con un valor general, para poder ir de lo general a lo especifico antes de llegar a cualquier conclusión.

Imagen que contiene Texto

Descripción generada automáticamente

Este paso lo lleve a cabo también para descartar realizar el conteo de población por duplicado, ya que, al realizar el análisis de la información sin ningún filtrado, note que el conteo era la suma de los valores por cada ciudad, adicionando también los valores totalizados que ya contemplaban los mismos. Esto me permitió con la información resultante, contabilizar cuantas veces se encontraba repetido cada condado en la lista y por otro lado sumar cuantas personas se encontraban en cada condado para conocer la población general, sin diferenciar el grupo etario o el sexo. Además, note que en la base de datos el conteo de condados de Irlanda era 42 en lugar de los 26 que corresponden, y esto se debe a que algunos, específicamente 16 casos, contemplaba la unión de condados que comparten frontera como uno adicional, esto para los casos de ciudades que se encuentran compartidas entre dos condados.

Una vez filtrados y conocidos los datos generales, aplique el método inverso, filtrando de la misma base de datos para excluir estos valores y obtener solo los específicos, estos contienen la diferenciación de sexo, grupo etario y ubicación en cada condado de la población, así como también el estado de salud de la población. Esto me ayudo a determinar en que condados se concentra la mayor densidad de población, que grupos etarios es el mas representativo y cuantos hombres y mujeres hay en la muestra. Utilizando estos valores comencé a seccionar la información, en una primera instancia dividí la distribución de hombres y mujeres y lo represente en un grafico de PIE, esto ya que es la mejor opción para mostrar esta distribución de manera gráfica, al ser solo dos variables, la muestra es clara y no permite errores en la interpretación de los valores.

Gráfico

Descripción generada automáticamente con confianza media

Mostrando unos valores de 51.34% (1.547.145) contra 48.66% (1.466.607), al ser una diferencia tan pequeña, procedi a aplicar uno de los métodos estudiados en el área de estadística, el método de Z-TEST, las mismas razones por las cuales aplique este test, son las razones por las cuales descarte utilizar un T-TEST, y lo explico a continuación:

Diagrama

Descripción generada automáticamente

Basado en este árbol de decisiones presentado en clases, al conocer la población estudiada ya queda descartada una de las ramas que nos recomienda utilizar el T-TEST, y por otro lado al ser una muestra representativa mayor a 30 registros queda descartado esta prueba, en favor del Z-TEST. Por otro lado, la razón por la que aplique este tipo de análisis es porque analizando la diferencia que existe entre la cantidad masculinos y femeninas podía determinar si este podía llegar a ser un factor en las diferencias que se obtuvieron en la comparación de grupos etarios o estados de salud general correspondiente a cada grupo. Según lo estudiado fije una hipótesis nula que para mi caso seria “La diferencia en sexo no es importante para el estudio”, teniendo en cuenta que el valor estándar utilizado para ALPHA en este tipo de pruebas es de 0.05, el valor de Z tendrá que ser mayor que 1,96 o menor que -1,96 para que esta hipótesis sea descartada, al realizar la prueba los resultados fueron los siguientes:

Texto

Descripción generada automáticamente

Por lo que -1.96 < -0,54 < 1,96, esto hace que la hipótesis nula sea aceptada y por esto determine que la diferencia en cantidad de hombres y mujeres estudiadas no es suficientemente representativa para que sea tomada en cuenta como un factor determinante en las siguientes fases de este estudio, debido a esto no lo tomo en cuenta como un factor que contribuya por ejemplo al deterioro de la salud en la base estudiada.

En mi siguiente paso conforme una tabla para clasificar los datos yendo a lo mas especifico y segmentado que se me ocurrió, teniendo el total de la información clasifique de la siguiente manera. Primero que estado de salud tiene la muestra, luego a que sexo corresponde, posteriormente a que grupo etario y por último a que condado. Basado en la información obtenida, mediante la variable County, realice graficas de barra para poder visualizar la cantidad de habitantes en cada condado organizados por grupo etario esto me permitió verificar en que grupos se concentra la población en cada condado para poder tomar la muestra mas representativa en los siguientes pasos en caso de necesitar centrar el estudio en una variable especifica que se pueda relacionar con las edades. A continuación, un ejemplo de los gráficos realizados:

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

De ahora en adelante en los siguientes pasos realizados procedí a relacionar las variables para segmentar y clarificar la información en distintos grupos por ejemplo en grupos etarios diferentes a fin de graficar la distribución, o por ejemplo para fijar un nivel de riesgo de las personas relacionando su edad con su estado de salud. Debido a que algunas de mis variables son nominales y otras ordinales, el paso lógico fue crear índices para mis variables a fin de poder relacionarlas y graficarlas sin problemas, ya que al intentar realizarlo en un primer momento tuve problemas al intentar relacionar los distintos STRINGS. Las primeras variables que codifique fueron los grupos etarios y los estados de salud que tengo en mi base de datos (esta codificación estará disponible en un archivo adicional de Excel que agregare como diccionario para mi base de datos).

A continuación, utilice un gráfico de barras para representar la cantidad de habitantes por cada grupo etario, utilizando un grafico de barras apiladas los segmente por su estado de salud, de esta manera obtuve una vista general de como se encuentra la distribución de estos, además obtuve el número total de personas contenidas en la base de datos por cada rango de edad, así como el porcentaje que representan del total de la población estudiada. Con esto determine que el 55,87% de la población se encuentra entre los 25 y los 59 años, y utilizando los gráficos antes realizados podemos observar por otro lado que es el grupo etario que goza de mejor estado de salud, por lo que reagrupe los grupos etarios a fin de estudiar esta población como una unidad:

Texto

Descripción generada automáticamente

Gráfico, Gráfico circular

Descripción generada automáticamente Además de obtener el número de individuos en cada grupo además grafique el porcentaje que representan en la totalidad:

En mi siguiente paso utilice la información codificada anteriormente para graficar esta información obtenida, relacione la cantidad de habitantes, en cada grupo etario y los clasifique por su estado de salud, y los represente en un gráfico de barrar, mediante esta visualización verificando la distribución en las categorías podemos visualizar que a mayor edad en la población, más dispersa es la distribución en los estados de salud así que puedo determinar que este es un factor influyente para el deterioro de la salud y es algo que tendré en cuenta en los siguientes pasos realizados en el estudio.

Una vez verificado mediante los gráficos que el grupo de mayores de 60 años es el que en promedio goza de una salud mas variada, considere que sería importante conocer la disposición geográfica de este grupo, esto lo hice determinando la cantidad de personas mayores de 60 años en cada condado, esto mediante la creación de la variable “seniors\_by\_county” y el conteo de los valores correspondientes a la muestra.

Ahora teniendo en cuenta que, bajo mi consideración a mayor edad, peor será el estado de salud, mayores serán los cuidades necesarios y mayor será el riesgo de muerte. Me pareció adecuado asignar un nivel de riesgo a cada estado de salud para determinar donde se ubica la población en esos niveles de “riesgo de muerte”, para realizar esta relación, me pareció adecuado definir un función a la cual llame “assign\_risk” la cual se aplico a todas las observaciones de la base de datos.

Texto

Descripción generada automáticamente

Mi siguiente paso fue graficar este riesgo para conocer la distribución de la población basado en esta característica, con esto podemos visualizar que la población censada en Irlanda goza de una buena salud general ya que la diferencia entre los valores de cada nivel de riesgo es bastante amplia en favor del nivel de salud “Very good”.

Gráfico

Descripción generada automáticamente

Aunque en un paso anterior determine que el sexo no es un factor determinante para el deterioro de la salud, me pareció interesante obtener el dato de cuantos masculinos y femeninas se encontraban contenidos en cada grupo de estado de salud, para poder realizar esta relación al igual que lo hice anteriormente codifique el sexo correspondiente a cada individuo y lo relacione con es estado de salud codificado para posteriormente con los valores poder graficarlo.

Luego de esto comencé a aplicar mis opciones para algoritmos de machine learnig, y para esto modifique nuevamente mi DF, codificando los niveles de riesgo a fin de contar con todas las variables en el mismo formato, y este nuevo DF lo llame “df\_encoded”

Tabla

Descripción generada automáticamente

Una vez con esta tabla, aplique un árbol de decisiones para determinar en que grupo de riesgo se deberían agrupar las observaciones en base a “Encoded Age Group” y 2Encoded General Health”. Debido a que tengo 3 variables y las mismas están bien determinadas puedo fijar ese numero como los niveles de mi árbol, además tomo como muestra el 50% de las observaciones en mi base de datos, destaco que al aplicar el árbol la precisión de este es del 100%, lo cual me deja saber la posibilidad de que el árbol este sobre ajustado, lo que lo haría sensible a errores o ruido en caso de valores que sean diferentes o no estén encuadrados en mis escalas. Aunque la conclusión a la que llego es que las variables que estoy referenciando en el árbol son bastante fijas y no dan lugar a interpretación, esto hace que por decirlo de una manera todas las posibilidades y combinaciones entre la salud y edad correspondan siempre en una única categoría.

Diagrama

Descripción generada automáticamente

La Explicación del árbol seria la siguiente:

Para una muestra del 50% de las observaciones (19620) el sistema se fija un umbral para la toma de decisiones, en el primer nivel este es 3.5 ya que basado en los estados de salud ese es el punto medio entre las dos categorías de riesgo (High Risk y Medium Risk), además tenemos los valores correspondientes a cada valor presente en la muestra. Otro valor presente en la caja es es el índice Gini, de manera sencilla, este índice me indica el nivel de homogeneidad o impureza del nodo, mientras mas bajo es este valor, en menos categorías está dividido el nodo.

Ya en el segundo nodo del árbol, podemos observar que se realizó la división y quedo de un lado los valores mayores o igual a 3.5 (los que representan buena salud y muy buena salud), con un Gini de 0.0 lo que nos indica que todos los valores en ese nodo pertenecen a una sola categoría y el nodo no se puede dividir nuevamente. Y por otro lado los valores que cumplían con la condición y eran menores o igual a 3.5, en este segundo nodo el Gini es de 0.44, esto nos indica que el nodo se podría dividir para generar nuevos nodos más homogéneos, así que fijando un nuevo umbral, en este caso <=1.5 que seria de nuevo el punto medio para el cambio de categoría entre Medium Risk y Low Risk se vuelve a dividir el nodo generando dos nuevas clasificaciones con un valor de Gini igual a cero lo que me indica que los nodos son completamente homogéneos y no pueden ser divididos nuevamente. Esto nos muestra una clasificación bastante clara de los valores presentes en la muestra.

Para mi segundo método de Machine Learning decidí usar Naïve Bayes. Anteriormente determine que las variables y salud están relacionadas, ya que a mayor edad, mas dispersa es la muestra en los estados de salud, aunque no están directamente relacionadas ya que no es una constante, por lo que para este caso la tomo como independientes, a fin de nuevamente ayudar con la clasificación en los niveles de riesgo, y para este caso nuevamente tome el 50% de la base para aplicar el algoritmo, relacionando los valores de nivel de salud y grupo etario se asigna un nivel de riesgo, para el caso de este método, la efectividad del sistema es del 100%, esto me sirve para dos cosas, una seria comprobar la pureza de mi base de datos y la segunda es sustentar los resultados obtenidos en mi árbol de decisiones.

Basado en los resultados de mi análisis de la información, teniendo en cuenta las graficas obtenidas y los resultados de los algoritmos de Machine Learning me atrevo a determinar que se podría utilizar la información en la base, por ejemplo para determinar en qué condados valdría la pena invertir en el área de la salud, basado en los niveles de riesgo de la población y en los rangos etarios, teniendo en cuenta la expectativa de vida de Irlanda, también se podría comenzar a diseñar in plan para pensar en fomentar una nueva distribución demográfica impulsando la movilización de la población a lugares donde el promedio de edad sea alto.

Para continual con los requerimientos como método de distribución discreta aplique la distribución geométrica para calcular la probabilidad de conseguir una persona con muy mala salud en un condado especifico, en este paso en el de mayor densidad demográfica (Counties Dublin & Meath). Obteniendo el siguiente resultado:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Este análisis me ayuda a determinar la posibilidad de conseguir una persona con muy mala salud en el condado especificado, teóricamente mide la posibilidad de tener éxito en un ensayo, teniendo variables con posibilidades constantes. Conocer este valor es importante porque nos ayuda a entender la magnitud de la cantidad de personas con ala salud presente en el condado, que para el caso aunque es el mas poblado, esto ayuda a que la cantidad de personas con baja salud sea poco representativa en el porcentaje total.