Estudio en la relación entre el área geográfica, el pueblo de pertenencia y el lugar de ocurrencia para fallecimientos entre 2009 a 2019

Axel López [1], Kevin Macario [2], Pablo Josue Nock Cajbon [3]

[1,2,3] Minería de datos, Departamento de Ciencias de computación y la información, Universidad del Valle de Guatemala

Resumen

En el presente trabajo se exponen evidencias sobre la relación que existe entre el servicio de salud que existe a lo largo del país de Guatemala y las muertes ocasionadas en los años 2009 - 2019.
Se pretende exponer si existe una relación entre el área geográfica, el pueblo de pertenencia y el sitio de ocurrencia del fallecimiento. Sin embargo se explorará las demás variables que existen en las bases de datos, ya que existen 30 variables, las cuales, 29 son cualitativas y 1 es cualitativa, a la cual fue la única variable a la que se le realizó un histograma.

Además, se explorará cada variable por separado, para identificar cualitativamente si tiene algún fenómeno que se debería explorar específicamente en la investigación . Posteriormente, al identificar las variables que sí comparten correlación y son importantes para la resolución de la hipótesis y objetivos específicamente planteados en la siguiente parte, las cuales se presentarán en el diagrama de correlación, se trabajará exclusivamente con ellas, y demostrara que estas comparten, o no, un coeficiente de incertidumbre entre ellas es siempre a 0.1.

Por último se discutirá sobre los hallazgos descubiertos y se interpretarán más a profundidad, con el fin de identificar cómo vive la sociedad guatemalteca en el ambiente de la salud, y si existe algún sesgo significativo entre el interior y exterior del país de Guatemala, ya que, como se sabe, en la actualidad está presente una centralización que puede afecta a los departamentos en los que no hay mucho desarrollo.

Introducción

Guatemala es uno de los países en donde la diferencia en el sector de salud entre la capital y el interior del país es bastante. Pese a que la mayoría de la población puede acceder al sistema de salud, muchos por distancia no pueden acceder y gran parte no reciben una atención de calidad. Esto eventualmente deriva en un alto porcentaje de fallecimientos sobre todo, como se mencionó con anterioridad. Al observar este comportamiento se puede llegar a deducir o a generar una hipótesis: “En Guatemala hay una relación entre el área geográfica, el pueblo de pertenencia y el sitio de ocurrencia del fallecimiento, de tal forma que el coeficiente de incertidumbre entre ellas es siempre a 0.1”.

Adicionalmente también como objetivos inspirados en las observaciones iniciales se refutan o confirmará si en Guatemala existe un sesgo debido al área geográfica, el pueblo de pertenencia y el sitio de ocurrencia del fallecido. Poder hacer uso de "Data mining" para sacar conclusiones de los resultados. Demostrar que mediante algoritmos de predicción se puede predecir con más de un 6% el área geográfica, el pueblo de pertenencia y el sitio de ocurrencia. Encontrar en qué medidas geográficas, el pueblo de pertenencia y el sitio de ocurrencia se encuentran relacionados.

Marco teórico

Situación problemática

En el Plan General de Gobierno de (Gobierno de Guatemala 2020–2024, 2020) citamos textualmente:

"El sector salud es otra área clave para alcanzar el desarrollo social. El país cuenta con una red hospitalaria compuesta por 46 hospitales, 281 centros de salud tipo B y 56 tipo A; así como, 916 puestos de salud, existiendo un alto nivel de concentración de los servicios de salud. El 73% de todos los médicos registrados y colegiados se encuentran en el departamento de Guatemala, lo que hace que la relación médica/población sea de 1 por cada 348 habitantes, mientras que en el interior de la República existe 1 médico por cada 11,489 habitantes.

El país también sufre los efectos de una atención deteriorada de la salud, como consecuencia del aumento poblacional, siendo las personas en condición de pobreza las más afectadas. La salud pública tiene una cobertura del 48% de la población, con un sistema de seguridad social que escasamente cubre el 16% de esa población."

Algoritmos empleados

Naive-Bayes

Son una clase especial de algoritmos de clasificación de Aprendizaje Automático. Se basan en una técnica de clasificación estadística llamada "teorema de Bayes".

Estos modelos son llamados algoritmos **Naive**, (Inocentes en español). En ellos se asume que las variables predictoras son independientes entre sí. Significa que la presencia de una cierta característica en un conjunto de datos no está relacionada con la presencia de cualquier otra característica.

Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Naive Bayes considera que cada una de estas características

contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad. Lo consiguen proporcionando una forma de calcular la probabilidad posterior de que ocurra un cierto evento A, dadas algunas probabilidades de eventos anteriores.

Dentro de sus características más destacables está que no es necesario una gran cantidad de datos de entrenamiento para que el algoritmo sea eficiente.

Este algoritmo lo podemos utilizar para predecir el número de muertes fetales por departamento, municipio, o región. Ya que se toman independiente cada dato y luego se predice la cantidad de muertes por lugar.

Regresión logística

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función de su estatura.

Una regresión lineal es un modelo estadístico que nos permite explicar los datos con una recta. Explica un modelo matemático (una recta) que relaciona la frecuencia cardíaca máxima con la edad. O dicho de otro modo, nos relaciona un fenómeno causa-efecto.

Podemos utilizar este algoritmo para la analizar los casos en que una madre que pierde a su hijo es casada o no. O también para saber el sexo de la persona fallecida duerante el año.

Neuronal Network:

La utilización de este algoritmo se debe a que una red neuronal nos permite clasificar de manera más rápida y eficiente gran cantidad de datos. Aunque son clasificadores muy precisos, no son comúnmente utilizadas para Data Mining porque producen modelos de aprendizaje inexplicables. En el presente trabajo se utilizará una clasificación por árboles de decisión por TREPAN, el cual a pesar de que algunas veces, los árboles de decisión extraídos no son lo suficientemente concisos. Para esos

casos se utilizan dos nuevos algoritmos de extracción de reglas que construyen árboles difusos y árboles modelo a partir de una Red Neuronal entrenada.

Decision tree

Este algoritmo trabaja como su nombre dice: tomando decisiones. Estas decisiones son binarias y son del tipo: ¿su sexo es masculino? si/no. En el caso de que sea así, entonces, procede a preguntar: ¿su edad es menor a los 8 años? si/no. Un ejemplo de esto, es la siguiente imagen, donde se trata el caso de las muertes del Titanic.

A partir de esto, podría pensarse al árbol de decisión como un algoritmo que aprende a hacer las preguntas indicadas para poder captar las características de las clases que pretende predecir, cómo se puede ver en la imagen anterior.

Ahora bien ¿cómo es que el árbol aprende a hacer las preguntas correctas?. Una manera sencilla en que esto se hace o se puede hacer es optimizar una función de costo. Las funciones de costo son simplemente funciones que miden un error entre el valor predecido y el valor real. Un ejemplo podría ser el coeficiente RMSE o algún error definido mediante entropías. Por lo tanto, un árbol de decisión empieza a proponer formas de ir clasificando nuestras características (o ir creando ramas) y luego mide la precisión (o error) que tuvimos mediante la función de costo y elegimos la clasificación que disminuya el error o aumente la precisión.

Un algoritmo para este tipo de clasificación, es el CART que es implementado en la librería de sklearn. Las ventajas de este algoritmo son:

- Es sencillo de interpretar
- Puede trabajar con variables numéricas, categorías y ambas.
- La no linealidad entre los parámetros no lo afecta.

Y sus desventajas son:

- Datos con mucha varianza pueden hacer que se generen árboles con decisiones completamente diferentes.
- Cuando existe sobreajuste, puede llegar a generar árboles demasiados complejos
- Se pueden generar árboles sesgados si existen clases mucho más comunes que otras.

Se va a emplear este algoritmo debido a que es bastante más sencillo que el Random Forest y requiere menos recursos (es una versión más sencilla de este) como una alternativa que pueda dar resultados precisos a un menor costo computacional. Otra razón es que dado a que el dataset tiene variables

categóricas, la matriz de correlación empleó entropía para medir la relación entre variables y este algoritmo tiene como función de costo a la entropía, lo que da más coherencia a la decisión de las variables que se empleen para predecir a nuestra variable objetivo.

Support Vector Machine:

Support Vector Machine es un algoritmo de aprendizaje supervisado. Está relacionado con problemas de clasificación y regresión.
Una de las grandes ventajas de este algoritmo es que trabaja eficientemente con grandes espacios de dimensiones y es muy útil con el uso de memoria, aunque cuando el número de variables es mayor al número de muestras puede ser que no sea tan eficiente en el uso de la memoria, produciendo que la implementación del mismo sea contraproducente.

Metodología

Fuente para la obtención de datos

Los datos se obtuvieron del Instituto Nacional de Estadística (INE), en la dirección: https://www.ine.gob.gt/ine/vitales/, la cual tiene los datos de las muertes de guatemaltecos en los períodos que comprenden desde enero de 2009 hasta diciembre de 2020.

Hipótesis

"En Guatemal hay una relación entre el área geográfica, el pueblo de pertenencia y el sitio de ocurrencia del fallecimiento, de tal forma que el coeficiente de incertidumbre entre ellas es siempre a 0.1"

Objetivos:

- **Objetivo Principal**
    - Encontrar que en Guatemala existe un sesgo debido al área geográfica, el pueblo de pertenencia y el sitio de ocurrencia del fallecido.
- **Objetivos específicos**
    - Poder implementar distintos algoritmos para trabajar con los datos
    - Demostrar que mediante algoritmos de predicción se puede predecir con más de un 6% el área geográfica, el pueblo de pertenencia y el sitio de ocurrencia.
    - Encontrar en qué medidas geográficas, el pueblo de pertenencia y el sitio de ocurrencia se encuentran relacionados.

Procesamiento de datos

Las bases de datos que se emplearon comprenden desde enero de 2009 hasta diciembre de 2019. A lo largo de los 10 años de datos se agregaron o quitaron variables al igual que se cambiaron el nombre de ciertas variables dado que representaban lo mismo solo la forma de denominar los datos cambiaron. Para los años que faltaban variables se crearon en las bases de datos esas variables con valores NA. Esto a manera de no perder información a la hora de crear una sola base de datos con los datos de todos los años.

Esto se hizo analizando cada base de datos por separado y comparando en R que variables tenía cada base de datos, se encontró que habían 29 variables cualitativas y 1 cuantitativa. Cabe resaltar que, cada base de datos tenia 854,681 registros y de todos estos areag, ocudif, escodif, pnadif, predif tenian numeroso registros con nan.

Habiendo hecho esto, las variables que conforman la bases de datos final son: departamento de registro del difunto "**depreg**", municipio de registro del difunto "**mupreg**", mes de registro "**mesreg**", año de registro "**añoocu**", departamento de ocurrencia "**depocu**", municipio de ocurrencia "**mupócu**", area geográfica "**areag**", sexo del difunto "**sexo**","**diaocu**","**mesocu**",edad del difunto "**edadif**", "**perdif**",pueblo del difunto "**puedif**", estado civil del difunto "**ecidif**",departamento de nacimiento del difunto "**dnadif**", municipio de nacimiento del difunto "**mnadif**","**nacdif**", departamento de residencia del difunto "**dredif**", municipio de residencia del difunto "**mredif**", causa de defunción "**caudef**",tipo de asistencia "**asist**", lugar de ocurrencia "**ocur**",certificado de defunción "**cerdef**", escolaridad del difunto "**escodif**", país de nacimiento del difunto "**pnadif**", "predfi", "**ciuodif**", "ocudif".

Cabe destacar que todas las variables son categóricas a excepción de la escolaridad, la edad del difunto y las variables referentes a tiempo (meses y años). Debido a la gran cantidad de variables categóricas, el valor de cada categoría se encuentra en la página de las estadísticas vitales del INE por cada año.

Matriz de correlación

En la matriz de correlaciones se aplicó la razón de correlación para variables categóricas-continuas y Theil's U para los casos categóricos-categóricos. A pesar de eso, el único dato que se considera continuo y no categórico son las edades de los difuntos, como vimos en la descripción de las variables. Además se clusterizaron las correlaciones para una fácil visualización. Cabe destacar que para todo esto se empleó el paquete : Dython para el lenguaje Python.

Se empleó Theil's U dado que esta es una medida asimétrica basada en entropía mutua, ya que no se vió en clase hay que aclarar entonces que esta medida predice que tanto mediante x podemos predecir y pero no que tanto podemos predecir y mediante x. En la tabla de correlaciones cada fila indica en qué proporción el índice describe a las otras variables, mientras que cada fila indica que tanto la variable es afectada por las otras.

Habiendo definido las variables de respuesta se definió un valor mínimo para considerar una variable regresora. el cual, es que cada variable tenga un coeficiente mayor o igual a 0.10, sin importar si es razón de correlación s o Theil's U. En el caso de la correlación se consideró que este puede ser positivo o negativo, por lo que el valor del coeficiente de correlación debe ser igual o mayor a abs(0.10). El valor para los coeficientes se eligió como 0.10 ya que se determinó que para todos los modelos de predicción, las variables regresivas que cumplen esta condición minimizan el error. Esto además se empleó para encontrar relaciones entre pares de variables y así graficar en tablas de frecuencias y otros gráficos para variables categóricas. Esto permitió definir cuales iban a ser las variables de respuesta, dado las relaciones o cómo era que estas parecían comportarse con las otras variables y encontrar relaciones no obvias o triviales.

Algoritmos empleados

Habiendo explorado pares de variables mediante sus valores en los coeficientes de la matriz de regresión y habiendo definido a partir de estos las variables regresivas, se procedió a la selección de los algoritmos a emplear. Los cuales fueron:

Naive-Bayes

Este algoritmo se seleccionó ya que con él se puede predecir el número de muertes fetales existentes por área, ya sea, departamento, municipio o región. La ventaja de este algoritmo es que todos los datos se toman como independientes y se posteriormente se predice la cantidad de muertes por lugar.

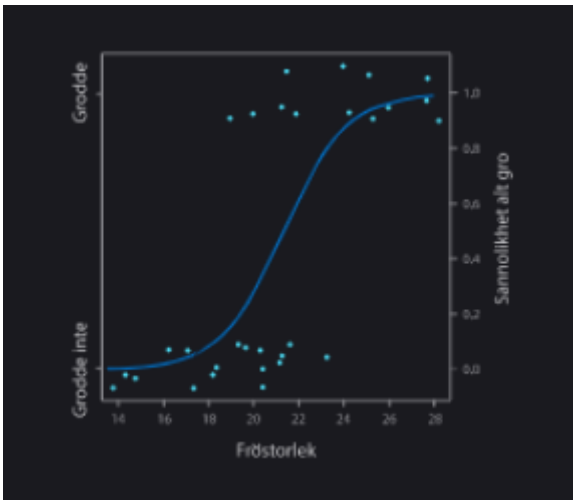
$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

- P(A): Probabilidad de A
- P(R|A): Probabilidad de que se de R dado A
- P(R): Probabilidad de R
- P(A|R): Probabilidad posterior de que se de A dado R

Regresión Logística

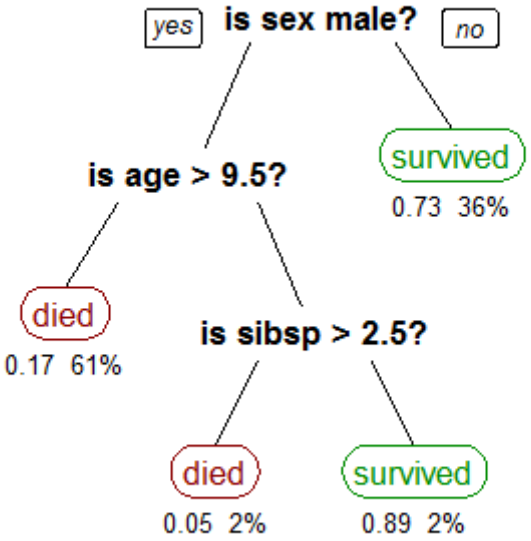
Se seleccionó este algoritmo, ya que se puede implementar por clasificación binaria en función de una variable cualitativa, y como se ha explicado anteriormente, las bases de datos seleccionadas tienen variables cualitativas en su mayoría, a comparación de variables cuantitativas.

Este algoritmo se utilizará para el análisis de casos, con las bases de datos, se clasificará cada coincidencia en donde una madre pierde a su hijo, posteriormente, si está casada o no. Ademas, tambien se clasificara el sexo de la persona fallecida durante el año, por mes.



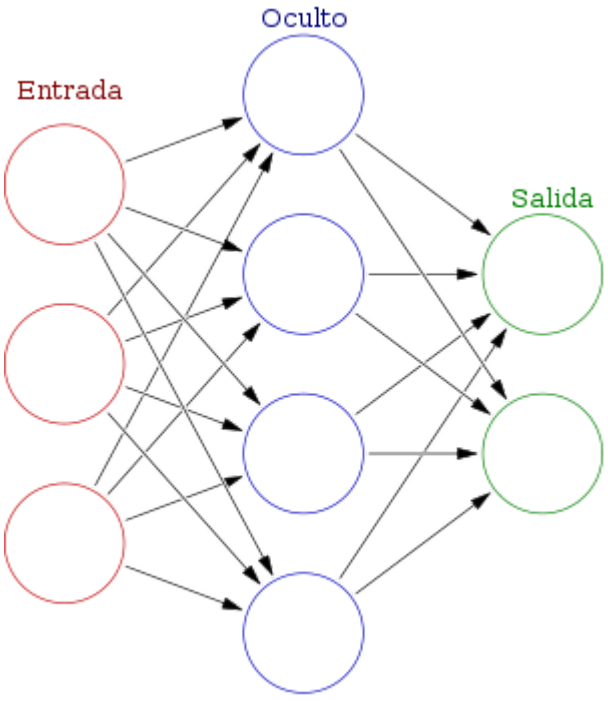
Decision tree

Se va a emplear este algoritmo debido a que es bastante más sencillo que el Random Forest y requiere menos recursos (es una versión más sencilla de este) como una alternativa que pueda dar resultados precisos a un menor costo computacional, de hecho, se determinó que era mejor usar árboles de decisión que random forest ya que dan resultados con precisiones parecidas y a un menor costo computacional. Otra razón es que dado a que el dataset tiene variables categóricas, la matriz de correlación empleó entropía para medir la relación entre variables y este algoritmo tiene como función de costo a la entropía, lo que da más coherencia a la decisión de las variables que se empleen para predecir a nuestra variable objetivo.



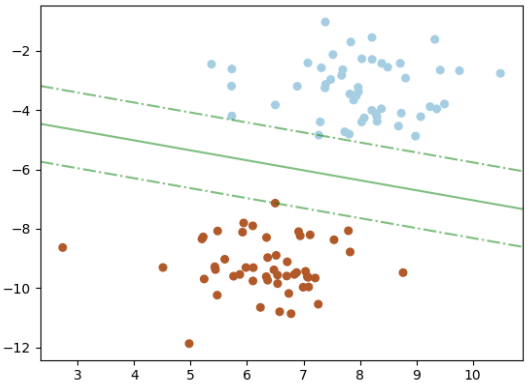
Neuronal Network:

A pesar de ser uno de los algoritmos más precisos, la precisión promedio que manejó el algoritmo fue del 78%, esto afectado en mayor medida porque algunas ocasiones faltan datos. Por lo que se recurrió a reemplazarlos por el valor 0, así dejando un valor reservado para los valor faltantes. Para nuestro caso en particular esto se corrigió de manera satisfactoria y no representó un problema en específico porque si llegó a clasificar los datos de manera precisa. Debido a que esto la precisión en las tres variables llegan a ser superiores al 78%.



Support Vector Machine:

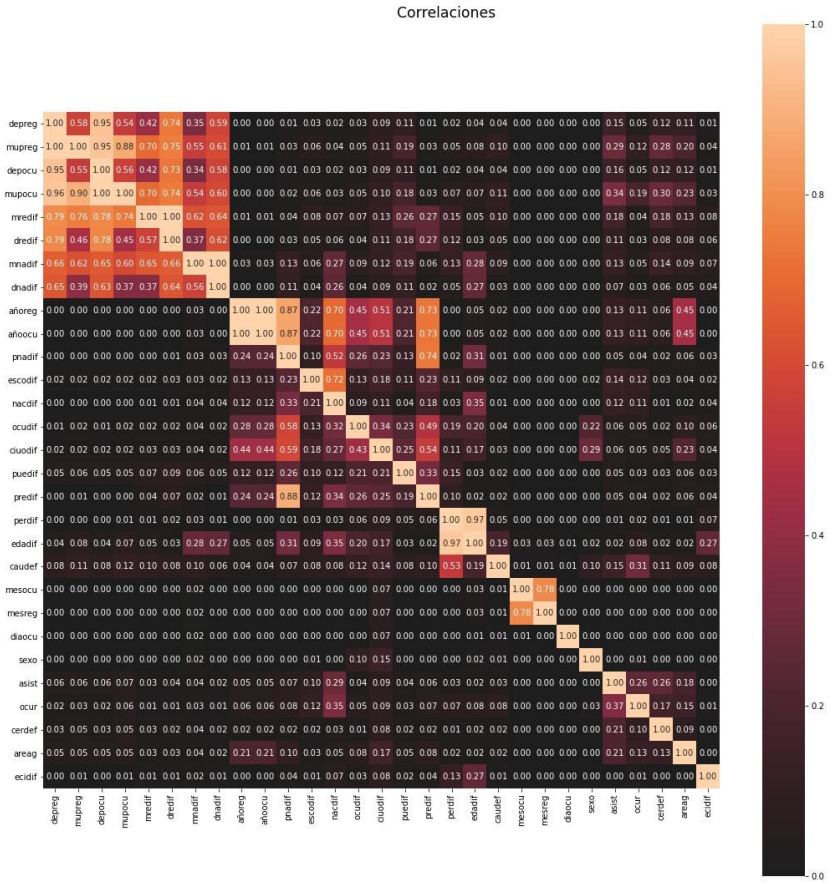
Se implementó este algoritmo ya que una de sus grandes ventajas es trabajar eficientemente con grandes espacios de dimensiones, y es muy bueno con el uso de memoria, sin embargo, cuando el número de características es mucho mayor al número de muestras puede que no sea tan eficiente



Por último es importante destacar que para todos los algoritmos se empleó validación cruzada y coeficiente R^2 para validar los resultados y su precisión, al igual que las matrices de confusión para poder profundizar más en como e intentar dar un porqué de los errores y aciertos de los algoritmos.

Discusión

Matriz de correlaciones



A partir de la matriz de correlación:

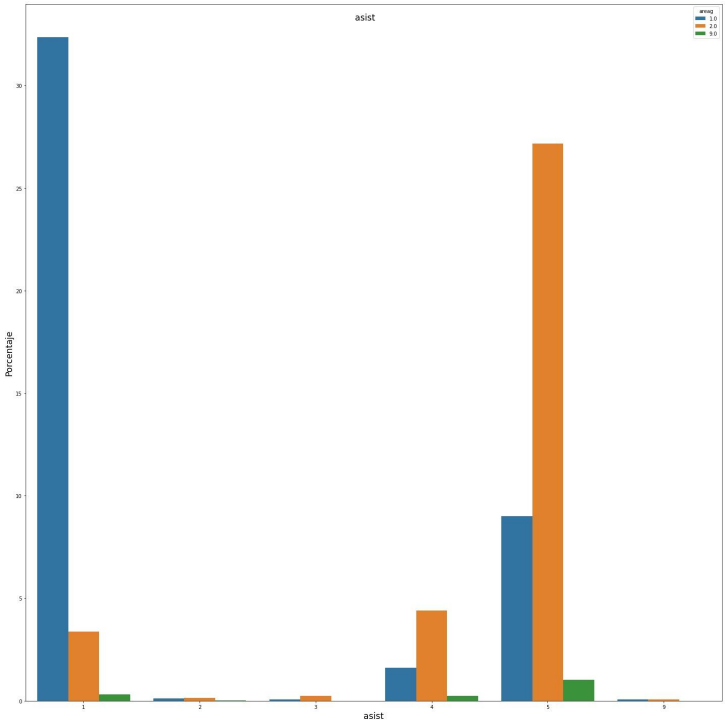
- Vemos que en general las variables depreg, mupreg, depocu, mupocu, mredif, dredrif, mnadif, dnadif se encuentran fuertemente relacionados entre ellos, como se evidenció en el análisis de variables individuales.
- El año de ocurrencia y el año de registro están también fuertemente relacionados con el país de nacimiento y la nacionalidad del difunto. Al igual que las variables ocudif y ciuodif, sin embargo, esto podría deberse a que el modelo responde a los NAs que hay en estas variables.
- Algo interesante a tomar en cuenta es que las variables ocudif y ciuodif parecen tener ambas una gran relación respecto al país de nacimiento del difunto.
- Parece ser que la escolaridad (escodif) no es descrita ni describe a otras variables como hubiese sido de esperarse, por ejemplo: la causa de defunción. Que no se encuentre correlación alta entre estas 2 variables implicaría que entre 2009 y 2019 en Guatemala la mayoría de muertes no tienen relación con la escolaridad de las personas.

- En el caso del sexo de las personas, vemos que su mayor relación es con ciuodif y ocudif (ocupaciones). Que se pueda describir la ocupación de las personas fallecidas en parte por su sexo, podría darse a cuestiones culturales.
- En general la causa de defunción está muy poco relacionada a las variables en general, excepto al país de residencia. Vemos que su capacidad de descripción de las variables es en general pequeña al igual que las variables que la describen a esta. Esto podría ser un indicio que para lograr un buen modelo de predicción es necesaria mucha información.
- Respecto a las edades vemos que esta describe a muchas más variables y su descripción también es recíproca, siendo la mayoría de variables a las que puede describir también las misma que la describen a esta variable.
- A partir del sitio de ocurrencia tenemos que hay cierta relación entre el lugar de nacimiento y el tipo de asistencia. Mientras que las variables que mejor la describen son también el tipo de asistencia y la causa de defunción y en menor medida otras variables relacionadas al año y al lugar de ocurrencia.

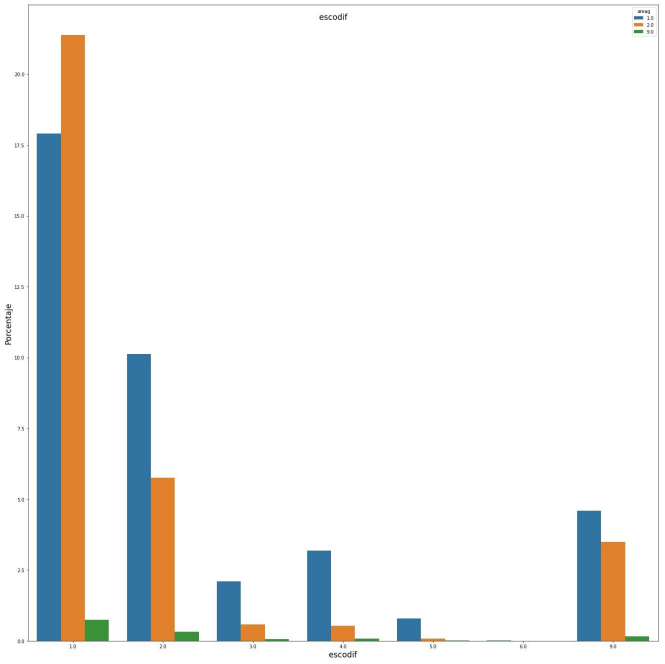
Análisis entre pares de variables

A partir de la matriz de correlación se buscaron las variables que más influyen a cada variable, que se representó como todas las variables que representarán en un 0.1 o más a las demás variables. Se puso dicho valor debido a que al disminuir el valor a 0.05, los modelos de predicción perdían precisión y al aumentarlo a 0.15, el modelo volvía a perder precisión, por lo tanto se asumió que 0.1 es aproximadamente el parámetro que optimiza a los modelos de predicción.Dado esto se encontraron los siguientes datos:.

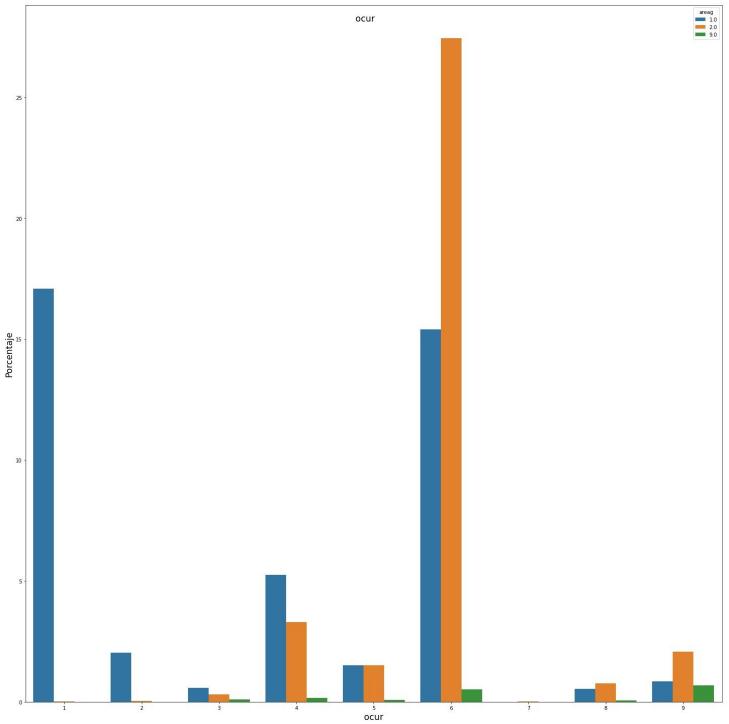
- areag:
    - Variables que la describen: ['depreg', 'mupreg', 'depocu', 'mupocu', 'mredif', 'añoreg', 'añoocu', 'ciuodif', 'asist', 'ocur']
    - Conclusiones interesantes: vemos que en general podemos saber el área geográfica del fallecido, esto podría ser importante ya que puede ser un indicativo de un sesgo en la atención que se brinda en las regiones. A partir de esto se hicieron gráficas de barras, de las cuales las conclusiones más interesantes son:
        - Areag - asist (tipo de asistencia):en el área urbana se recibe mayormente ayuda médica, sin embargo, en el área rural predomina la ausencia de esta y en menor medida se recibe ayuda empírica.



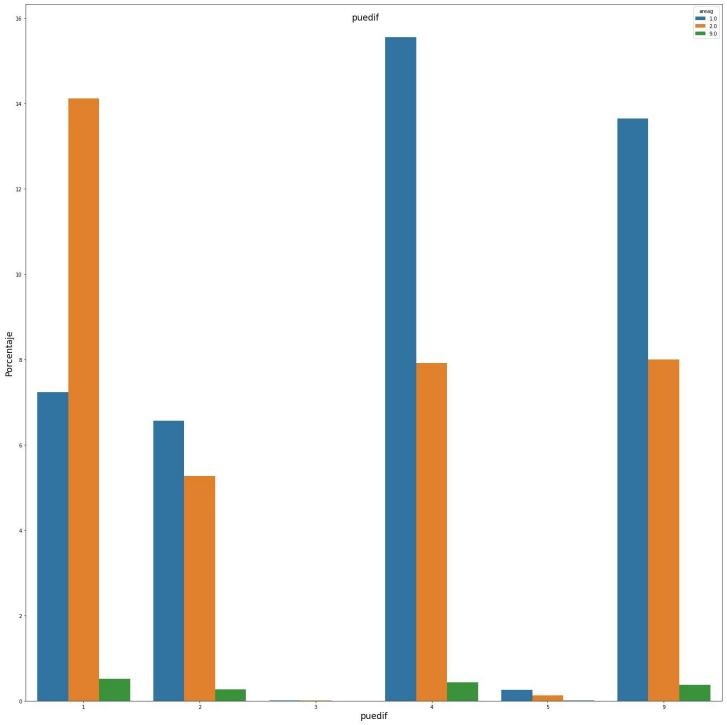
- Areag-escodif(escolaridad del difunto): tenemos que la mayoría de personas que fallecieron en el área rural no tenían ningún tipo de educación.También vemos que en proporción, esta frecuencia disminuye más rápidamente conforme el grado de escolaridad aumenta. Y vuelve a aumentar para los casos en que se ignora el grado de escolaridad de la persona.



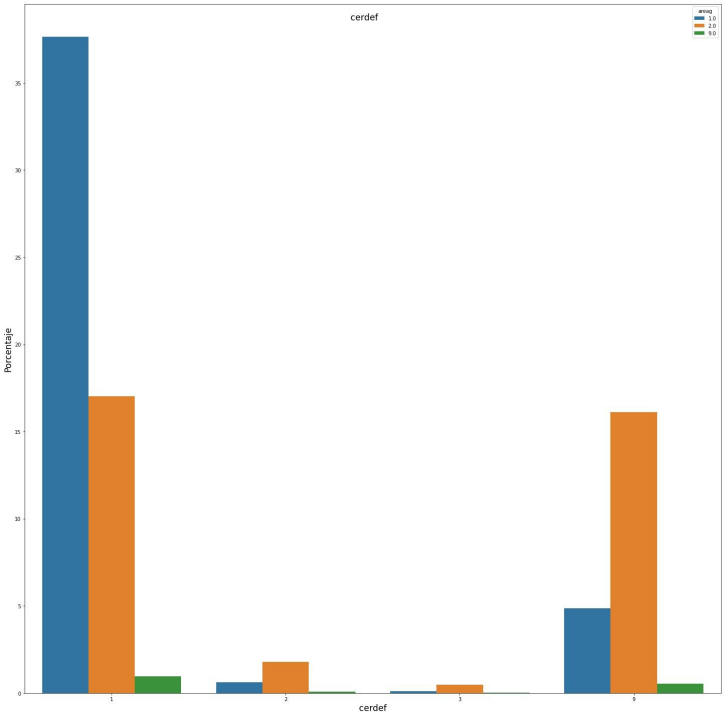
- Areag-Ocur (sitio de ocurrencia): la mayoría de veces los fallecimientos que ocurren en el área urbana suceden en hospitales, esto contrasta con la distribución de fallecimientos en domicilios en el área rural. También cabe destacar que la cantidad de fallecimientos en domicilios en el área urbana es similar a los que ocurren en hospitales.Por último se observa que la cantidad de fallecimientos en la vía pública tanto en el área rural y urbana son similares y que hay más fallecimientos en sitios desconocidos en el área rural que el área urbana.



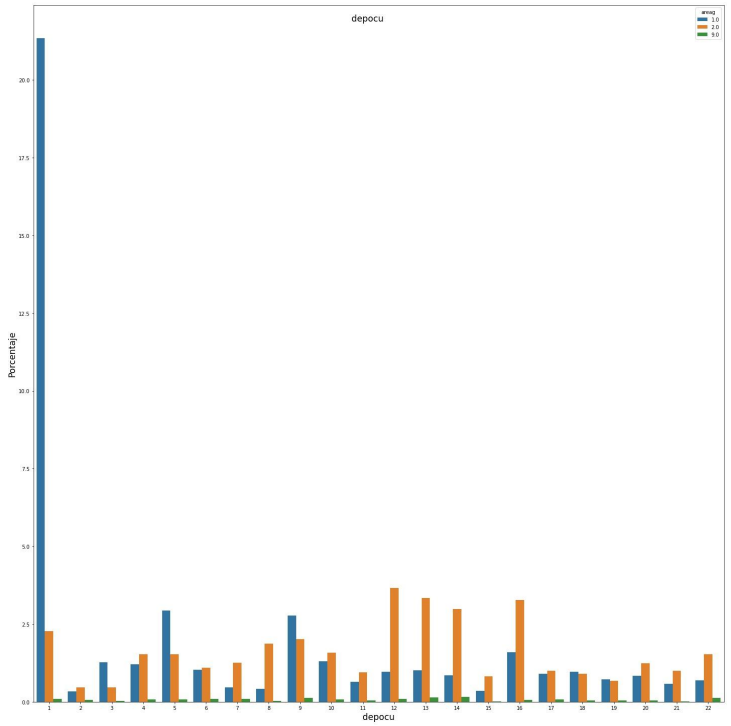
- Areag-puedif(Pueblo del difunto): el pueblo al que pertenecen la mayoría de personas que fallecen en el área rural es el pueblo Maya.



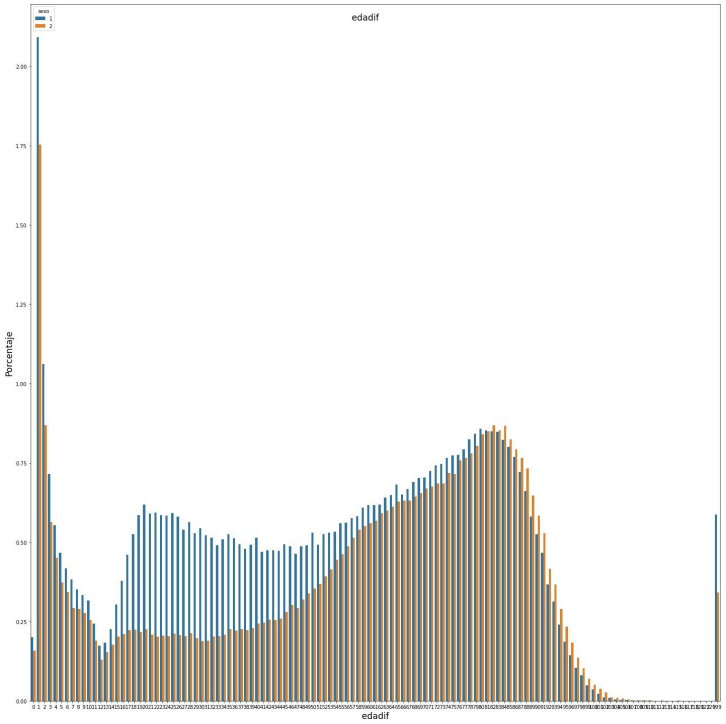
- Areag-cerdef (Quien certifica): la proporción de personas que son certificadas por un médico es mayor en el área urbana, mientras que la proporción de personas que se es ignorado quién certifica son del área rural.



- Areag-Depocu (departamento de ocurrencia): en el departamento de Guatemala hay una mayor proporción de personas que fallecen en el área urbana, sin embargo, existen departamentos donde la mayoría de fallecimientos se dieron en el área rurall. En general esto se puede generalizar para los departamentos y municipios de ocurrencia, registro, nacimiento y residencia.



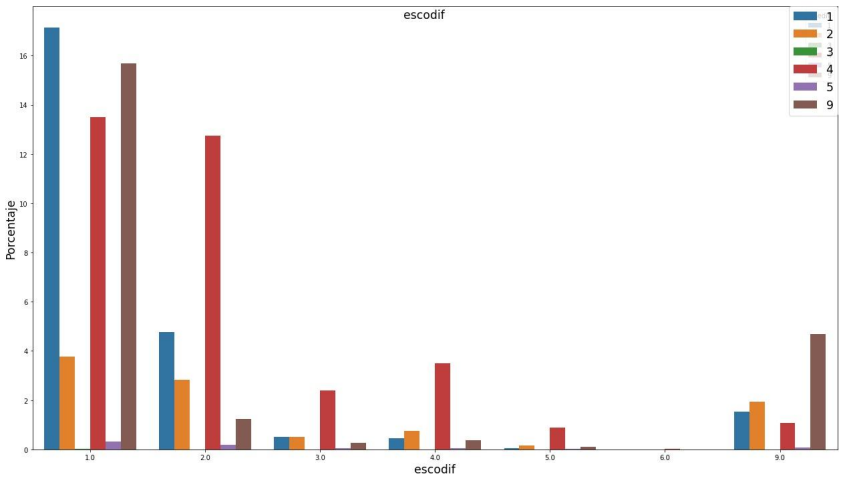
- sexo:
  - Variables que la describen: ['ocudif', 'ciuodif']
  - Conlusiones interesantes: en general las gráficas de barras demuestran que para cualquier agrupación la cantidad de muertes es mayor en hombres que para mujeres. Sin embargo, se encontró que para las edades de los difuntos (edadif), existe una diferencia en la frecuencia de cada clase para hombres y mujeres, generando una distribución de datos que para los extremos convergen pero para valores intermedios cambian drásticamente. A partir de esto, se puede concluir que en Guatemala es menos probable que una mujer fallezca en su juventud y parte de su edad adulta. Pero conforme envejece su probabilidad de fallecer es la misma que para los hombres.



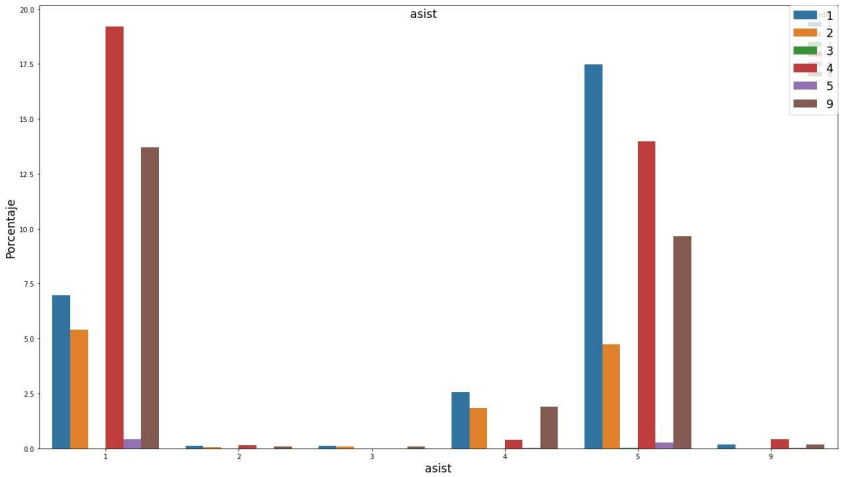
- Conclusiones interesantes: para esta variable es de esperarse que sea descrita por la causa de defunción, ocupación y el estado civil, sin embargo, es curioso que la edad a la que las personas fallecen también esté relacionado con el municipio de nacimiento y el departamento, esto podría implicar que a tempranas edades los niños fallecen más en un municipio y departamentos que en otros.

- puedif:
  - Variables que la describen: ['depreg', 'mupreg', 'depocu', 'mupocu', 'mredif', 'dredif', 'mnadif', 'dnadif', 'añoreg', 'añoocu', 'pnadif', 'escodif', 'ocudif', 'ciuodif', 'predif']
  - Conclusiones interesantes: vemos que esta variable está descrita por variables referentes a los lugares en donde se registró, dónde ocurrió, la escolaridad y la profesión de cada pueblo. Esto podría deberse a que la distribución de pueblo en Guatemala no es uniforme. A partir de esto se hicieron gráficas de barras, de las cuales las conclusiones más interesantes son:
    - Puedif - Escodif (escolaridad del difunto): en general parece ser que fallecen más personas del pueblo maya y ladino que los garífunas y Xincas. Sin embargo, esto podría deberse al número de personas que pertenecen a cada pueblo. Algo a remarcar es que en proporción mueren más personas personas indígenas sin estudios que a comparación de otros pueblos, sin embargo, sin importar el nivel de educación la proporción cambia y fallecen más personas pertenecientes al pueblo ladino. Esto podría deberse a distintos factores, uno

de estos, podría ser al acceso de la educación que tiene cada pueblo. También vemos que la cantidad de personas a las que no se conoce su nivel educativo está relacionado con no saber de qué pueblo procede el fallecido.



- Puedif - asist (asistencia recibida): el pueblo ladino es el que más se le da asistencia médica, sin embargo, algo curioso es que también se le dió a una gran cantidad de personas de las cuales se desconoce el pueblo al que pertenecen. En el caso de la ausencia de atención médica, vemos que la mayoría de casos fueron con el pueblo Maya, mientras que en segundo lugar se encuentra el pueblo ladino y también se tiene una gran proporción de personas a las que no se conoció el pueblo de procedencia.



Algoritmos de predicción

Dando que ciertas variables tienen muchos nan, lo que se hizo fue utilizar de nuevo las variables que mejor describen a la variable que vamos a estudiar mediante sus coeficientes de la matriz de correlación. Sin embargo, si esta era descrita por variables con muchos nan, estas se descartaron a la hora de emplear el modelo de regresión, a excepción de algunas variables que se eligieron empíricamente pero que siempre cumplía con tener un coeficiente mayor o igual a 0.1.

Por otro lado, las variables que se van a predecir serán: areag,escodif, puedif y ocur. La relación entre la la alta mortalidad en niños y su relación con el municipio de nacimiento parecían ser relaciones interesantes a estudiar, sin embargo, se considera que hacen faltas otras métricas, como la cantidad de nacimientos, la población o pirámides de población.

Los algoritmos que se utilizaron fueron:

- Regresión logística
    - Parámetros modificados:
        - fit_intercept = True
        - solver = 'lbfgs'
        - random_state = 666
        - warm_start = True
        - penalty = 'l2'
        - n_jobs = -1)

- árboles de decisión:
    - Parámetros modificados
        - criterion='entropy',
        - min_samples_split=20,
        - min_samples_leaf=5,
        - random_state = 666)

- Naive Bayes
    - Parámetros modificados
        - Ninguno

Se usaron estos algoritmos en general porque se sabe que son bastante buenos para la clasificación, además, la mayoría fueron vistos en clase.

Para los métodos de validación se utilizó validación cruzada y la función score (R^2) del scikit learn de cada algoritmo tanto en los sets de entrenamiento y prueba y se generaron matrices de confusión para los datos de prueba.

Precisión y matrices de correlación

Área geográfica

- Regresión logística
  - Coeficientes

Forma de validación	Valor
Validación cruzada	0.75
R^2 train	0.7814869184971166
R^2 test	0.7822254065167279

  - Matriz de confusión

```
no normalizado
Prediccion    1.0    2.0  9.0    All
Verdadero
1.0         28024   6144  265  34433
2.0          6228  22188   70  28486
9.0           322    953   10   1285
All         34574  29285  345  64204

***************

normalizado
Prediccion       1.0       2.0       9.0
Verdadero
1.0         0.436484  0.095695  0.004127
2.0         0.097003  0.345586  0.001090
9.0         0.005015  0.014843  0.000156
```

Vemos que para el área geográfica la validación cruzada da un 0.75% de precisión para el modelo. Esto se contrasta con el coeficiente R^2 del set de entrenamiento y de prueba los cuales dan un valor ligeramente más alto, además, de estos 2 coeficientes vemos que no hay sobre ajuste y si lo hay es muy pequeño.

Respecto a las tablas de confusión vemos que el error entre la predicción del suceso en el área rural y el área urbana es parecido y en general sus aciertos también lo son, sin embargo, para el valor

desconocido el error es alto (es donde más se confunde), esto podría deberse a 2 condiciones principales: hay muy pocos datos en los que se desconozca el lugar del deceso, o bien, el modelo no fué capaz de encontrar una característica con la que pueda definir a los decesos y diferenciarlo de los otros 2 casos, por lo que, empleó las características del área rural y urbana para clasificarlo

- Árboles de decisión
    - Coeficientes

Forma de validación	Valor
Validación cruzada	0.85
R^2 train	0.8739978116372613
R^2 test	0.84923057753411

    - Matriz de confusión

```
no normalizado
Prediccion    1.0    2.0  9.0    All
Verdadero
1.0         28024   6144  265  34433
2.0          6228  22188   70  28486
9.0           322    953   10   1285
All         34574  29285  345  64204

***************

normalizado
Prediccion       1.0       2.0       9.0
Verdadero
1.0         0.436484  0.095695  0.004127
2.0         0.097003  0.345586  0.001090
9.0         0.005015  0.014843  0.000156
```

Vemos que para el área geográfica la validación cruzada da un 0.85% de precisión para el modelo. Esto se contrasta con el coeficiente R^2 del set de entrenamiento y de prueba los cuales dan un valor ligeramente más alto, además, de estos 2 coeficientes vemos que hay un sobreajuste, pero este es

pequeño siendo de aproximadamente el 2% la diferencia entre la precisión de los datos de entrenamiento y prueba.

Respecto a la matriz de confusión vemos que entre al área rural y urbana se mantiene la misma precisión a la hora de predecirlas, siendo más preciso para el área urbana que para el área rural, también vemos que para los datos desconocidos esta vez el modelo fue capaz de predecir los mejor, aunque ahora vemos que hay una gran cantidad de desconocidos que el modelo los tomó como área rural

- Naive Bayes
    - Coeficientes

Forma de validación	Valor
Validación cruzada	0.83
R^2 train	0.83
R^2 test	0.83

    - Matriz de confusión

```
no normalizado
Prediccion    1.0    2.0   9.0    All
Verdadero
1.0         28481   5623   329  34433
2.0          3104  24637   745  28486
9.0           217    727   341   1285
All         31802  30987  1415  64204

***************

normalizado
Prediccion       1.0       2.0       9.0       All
Verdadero
1.0         0.443602  0.087580  0.005124  0.536306
2.0         0.048346  0.383730  0.011604  0.443680
9.0         0.003380  0.011323  0.005311  0.020014
All         0.495327  0.482633  0.022039  1.000000
```

Vemos que para el área geográfica la validación cruzada da un 0.83% de precisión para el modelo. Esto se contrasta con el coeficiente R^2 del set de entrenamiento y de prueba los cuales obtuvieron el mismo valor, además, por lo que se asume que el sobre ajuste es mínimo.

Respecto a la matriz de confusión vemos que entre al área rural y urbana se mantiene la misma precisión a la hora de predecirlas, siendo más preciso para el área urbana que para el área rural, también vemos que para los datos desconocidos esta vez el modelo fue capaz de predecir los mejor que con la regresión logística y parecido al árbol de entrenamiento , aunque ahora vemos que hay una gran cantidad de desconocidos que el modelo los tomó como área rural y muchos del área rural que se tomaron como desconocidos. Esto como se discutió anteriormente, podría ser un indicador de que realmente la mayoría de valores que se tomaron como desconocidos fueron realmente en el área rural.

<u>Pueblo de pertenencias</u>

- Regresión logística
    - Coeficientes

Forma de validación	Valor
Validación cruzada	0.46
R^2 train	0.46
R^2 test	0.46

- Matriz de confusión

```
no normalizado
Prediccion      1    2  3      4      9    All
Verdadero
1            8359   26  0   9991   3544  21920
2            4976   43  5     10   4597   9631
3               8    0  0     10      0     18
4            5462   84  0  18222   3600  27368
5              85    2  0    424     64    575
9            5041   41  4   3222  12180  20488
All         23931  196  9  31879  23985  80000

***************

normalizado
Prediccion         1         2         3         4         9       All
Verdadero
1           0.104487  0.000325  0.000000  0.124887  0.044300  0.274000
2           0.062200  0.000538  0.000063  0.000125  0.057462  0.120387
3           0.000100  0.000000  0.000000  0.000125  0.000000  0.000225
4           0.068275  0.001050  0.000000  0.227775  0.045000  0.342100
5           0.001063  0.000025  0.000000  0.005300  0.000800  0.007188
9           0.063012  0.000513  0.000050  0.040275  0.152250  0.256100
All         0.299138  0.002450  0.000112  0.398487  0.299812  1.000000
```

Vemos que en general el modelo obtuvo un desempeño muy pobre a la hora de predecir el pueblo al que los difuntos eran originarios

- Árboles de decisión
  - Coeficientes

Forma de validación	Valor
Validación cruzada	0.78
R^2 train	0.81
R^2 test	0.79

- Matriz de confusión

no normalizado

Prediccion	1	2	4	5	9	All
Verdadero						
1	18600	455	1434	12	1419	21920
2	493	6922	13	0	2203	9631
3	0	0	18	0	0	18
4	1874	1	24564	24	905	27368
5	172	0	295	32	76	575
9	2652	2014	3288	25	12509	20488
All	23791	9392	29612	93	17112	80000

****

normalizado

Prediccion	1	2	4	5	9	All
Verdadero						
1	0.232500	0.005687	0.017925	0.000150	0.017737	0.274000
2	0.006162	0.086525	0.000162	0.000000	0.027537	0.120387
3	0.000000	0.000000	0.000225	0.000000	0.000000	0.000225
4	0.023425	0.000013	0.307050	0.000300	0.011312	0.342100
5	0.002150	0.000000	0.003687	0.000400	0.000950	0.007188
9	0.033150	0.025175	0.041100	0.000313	0.156362	0.256100
All	0.297388	0.117400	0.370150	0.001162	0.213900	1.000000

Vemos que el árbol de clasificación se desempeñó bastante bien, dado que tanto la precisión en la validación cruzada y los coeficientes R2 para los set de entrenamiento y prueba predijeron 30% mejor todas las clases que la regresión logística.

Con respecto a la matriz de correlación podemos observar que en general los resultados se asemejan con las demás variables, teniendo el modelo dificultad para predecir correctamente los casos donde son desconocidos los pueblos de proveniencia. Pero, a pesar de esto, vemos que acertó una cantidad mucho más grande de desconocidos que en otros casos, esto puede ser un indicativo que en esta variable hay razones mucho más claras del porqué se desconoce el pueblo de proveniencia.

- Naive Bayes

- Coeficientes

Forma de validación	Valor
Validación cruzada	nan
R^2 train	0.71
R^2 test	0.71

- Matriz de confusión

no normalizado

Prediccion	1	2	4	5	9	All
Verdadero						
1	16624	2176	1428	290	1402	21920
2	1777	6484	19	1	1350	9631
3	0	0	18	0	0	18
4	1982	0	24455	354	577	27368
5	150	0	278	102	45	575
9	3037	3777	3680	189	9805	20488
All	23570	12437	29878	936	13179	80000

****

normalizado

Prediccion	1	2	4	5	9	All
Verdadero						
1	0.207800	0.027200	0.017850	0.003625	0.017525	0.274000
2	0.022212	0.081050	0.000237	0.000013	0.016875	0.120387
3	0.000000	0.000000	0.000225	0.000000	0.000000	0.000225
4	0.024775	0.000000	0.305688	0.004425	0.007213	0.342100
5	0.001875	0.000000	0.003475	0.001275	0.000562	0.007188
9	0.037963	0.047212	0.046000	0.002363	0.122563	0.256100
All	0.294625	0.155463	0.373475	0.011700	0.164738	1.000000

El resultado de Naive-Bayes no difiere mucho respecto al del árbol de clasificación, aunque ahora, la precisión disminuyó , a pesar de eso vemos que (aunque con más error) es capaz de predecir la mayoría de clases, con la diferencia de que ahora la clase de desconocido tiene errores mucho más grandes.

Lugar de ocurrencia

- Regresión logística
    - Coeficientes

Forma de validación	Valor
Validación cruzada	0.60
R^2 train	0.59
R^2 test	0.59

    - Matriz de confusión

no normalizado

Prediccion	1	2	3	4	6	All
Verdadero						
1	2591	15	473	720	9673	13472
2	149	3	4	26	1446	1628
3	32	0	280	294	149	755
4	524	2	220	1960	4285	6991
5	1411	0	0	1	1079	2491
6	1535	16	0	3	33341	34895
7	8	0	0	0	5	13
8	439	0	0	1	622	1062
9	636	1	11	24	2225	2897
All	7325	37	988	3029	52825	64204

****

normalizado

Prediccion	1	2	3	4	6	All
Verdadero						
1	0.040356	0.000234	0.007367	0.011214	0.150660	0.209831
2	0.002321	0.000047	0.000062	0.000405	0.022522	0.025357
3	0.000498	0.000000	0.004361	0.004579	0.002321	0.011759
4	0.008161	0.000031	0.003427	0.030528	0.066740	0.108887
5	0.021977	0.000000	0.000000	0.000016	0.016806	0.038798
6	0.023908	0.000249	0.000000	0.000047	0.519298	0.543502
7	0.000125	0.000000	0.000000	0.000000	0.000078	0.000202
8	0.006838	0.000000	0.000000	0.000016	0.009688	0.016541
9	0.009906	0.000016	0.000171	0.000374	0.034655	0.045122
All	0.114089	0.000576	0.015388	0.047178	0.822768	1.000000

Vemos que el modelo obtuvo entre 59% y 60% en la precisión de sus predicciones, sin embargo, este no fue capaz de predecir todas las categorías de la variable a estudiar.

Árboles de decisión

- Coeficientes

Forma de validación	Valor
Validación cruzada	0.84
R^2 train	0.86
R^2 test	0.85

- Matriz de confusión

no normalizado

Prediccion	1	2	3	4	5	6	7	8	9	All
Verdadero										
1	11510	135	29	589	22	1155	0	13	19	13472
2	775	111	10	129	9	580	0	6	8	1628
3	71	8	527	86	7	40	0	1	15	755
4	1701	43	118	4432	10	663	0	0	24	6991
5	57	3	1	3	1995	151	0	97	184	2491
6	1272	116	14	164	113	32805	1	34	376	34895
7	1	0	0	0	4	4	0	3	1	13
8	54	4	0	4	160	220	0	469	151	1062
9	56	3	25	9	201	389	0	100	2114	2897
All	15497	423	724	5416	2521	36007	1	723	2892	64204

****

normalizado

Prediccion	1	2	3	4	5	6	\
Verdadero							
1	0.179272	0.002103	0.000452	0.009174	0.000343	0.017990	
2	0.012071	0.001729	0.000156	0.002009	0.000140	0.009034	
3	0.001106	0.000125	0.008208	0.001339	0.000109	0.000623	
4	0.026494	0.000670	0.001838	0.069030	0.000156	0.010326	
5	0.000888	0.000047	0.000016	0.000047	0.031073	0.002352	
6	0.019812	0.001807	0.000218	0.002554	0.001760	0.510949	
7	0.000016	0.000000	0.000000	0.000000	0.000062	0.000062	
8	0.000841	0.000062	0.000000	0.000062	0.002492	0.003427	
9	0.000872	0.000047	0.000389	0.000140	0.003131	0.006059	
All	0.241371	0.006588	0.011277	0.084356	0.039265	0.560822	

Prediccion	7	8	9	All
Verdadero				
1	0.000000	0.000202	0.000296	0.209831
2	0.000000	0.000093	0.000125	0.025357
3	0.000000	0.000016	0.000234	0.011759
4	0.000000	0.000000	0.000374	0.108887
5	0.000000	0.001511	0.002866	0.038798
6	0.000016	0.000530	0.005856	0.543502
7	0.000000	0.000047	0.000016	0.000202
8	0.000000	0.007305	0.002352	0.016541
9	0.000000	0.001558	0.032926	0.045122
All	0.000016	0.011261	0.045044	1.000000

Aunque existe un poco de sobreajuste, dada la matriz de correlación se considera que el modelo es bastante bueno para predecir las categorías teniendo un 84 % en la validación cruzada y un 86% y 85% con los datos de entrenamiento y prueba respectivamente.

Además en las matrices de confusión vemos que en general pudo predecir cada categoría bastante bien, incluyendo los desconocidos.

Support Vector Machine

Se realizó Support Vector Machine con tres diferentes Kernels (lineal, polinomial y RBF) utilizando las variables: areaag, muspreg, ocudif y depreg. Se eligieron estas variables ya que eran variables que compartían bastante correlacionalidad, de acuerdo a la gráfica de correlación que se realizó.

Luego de implementar el svm se realizó un cross validation para identificar qué tan independientes eran los datos de prueba y de testeo, lo cual nos resultó en un arreglo de porcentajes que determinaban qué tan preciso era el modelo, para trabajar con estos se hizo un promedio de todo el arreglo.
Sin embargo, la ejecución de este algoritmo resultaba demasiado ineficiente ya que tardaba mucho la ejecución, y los scores que retornaba el algoritmo no eran buenos, ya que no superan el 0.7
Cabe resaltar que este procedimiento se hizo con cada uno de las variables y con cada uno de los tres kernels.

La pobre ejecución de SVM se puede deber a que las variables pueden no estar tan correlacionadas como se planteó al principio de la investigación. Esto da paso a rechazar la hipótesis, tomando en cuenta los 4 algoritmos implementados anteriormente y el actual.

Conclusiones

Se rechaza la hipótesis debido a que se demostró con los algoritmos utilizados que no hay una relación directa entre el área geográfica, pueblo de pertenencia y el sitio del fallecimiento, puesto que el coeficiente de incertidumbre entre ellas siempre supera 0.1

Como se demostró en las gráficas de agrupamiento existe un cambio significativo en las condiciones en las que se atendió a la persona fallecida, se identificó que este esta con función al área geográfica, pueblo de pertenencia y sitio de ocurrencia.

Se logró cumplir el primer objetivo específico ya que fue posible predecir de manera eficaz las categorías con más de 70% de precisión.

Se cumplió el segundo objetivo ya que se encontraron coeficientes de indeterminación entre las variables.

Se encontró que el pueblo ladino fallece mas en hospitales, y el pueblo maya fallece mas en su casa, esto se puede deber a que en el interior del país carecen de asistencia médica a comparación de el departamento de Guatemala.

Referencias

Gobierno de Guatemala 2020–2024. (2020). Problemática de País: Desarrollo Social | Vicepresidencia de la República de Guatemala. Recuperado 5 de abril de 2021, de https://vicepresidencia.gob.gt/politica-gobierno-2020-2024/Problematica-de-Pais-Desarrollo-Social

Zychlinski, S. (2019, 26 diciembre). The Search for Categorical Correlation - Towards Data Science. Recuperado de https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9
Amat, R. (2016) Regresión logística simple y múltiple. Recuperado de: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

Roman, V. (2019) Algoritmos Naive Bayes: Fundamentos e Implementación. Recuperado de: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementación-4bcb24b307f

Yiu, T. (2021, March 28). Understanding Random Forest - Towards Data Science. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Gupta, P. (2018, June 20). Decision Trees in Machine Learning - Towards Data Science. Medium. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052