Universidad del Valle de Guatemala
Minería de Datos
Lynette García

Presentación de resultados

Mario Sarmientos - 17055
Joonho Kim - 18096
Augusto Alonso - 181085

Explicación de Métodos utilizado

Random Forest

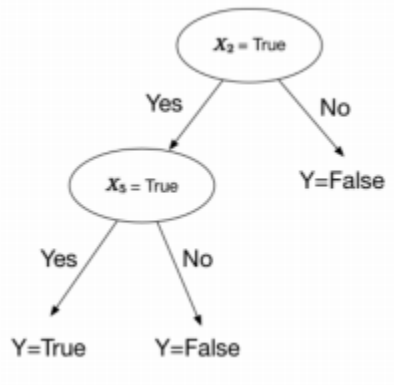
Random forest En este algoritmo se crean varios árboles, para clasificar un nuevo objeto basado en atributos. Cada árbol da una clasificación y finalmente el bosque elige la clasificación con más votos. Obviamente cuanto más árboles haya en el bosque más robusto será el bosque y por ende la precisión del algoritmo. Se puede ver en la gráfica de manera más práctica, tenemos una data la cual la dividimos en tres para crear tres árboles de decisión. Al final, cada árbol arroja un resultado y la respuesta que tenga mayor voto, será la predicción de nuestro análisis. Al igual que todos los problemas de aprendizaje supervisado siempre tenemos que tener las variables independientes y dependientes. En donde la variable independiente o características, es la que se manipula para determinar el valor de una variable dependiente.

Decision tree

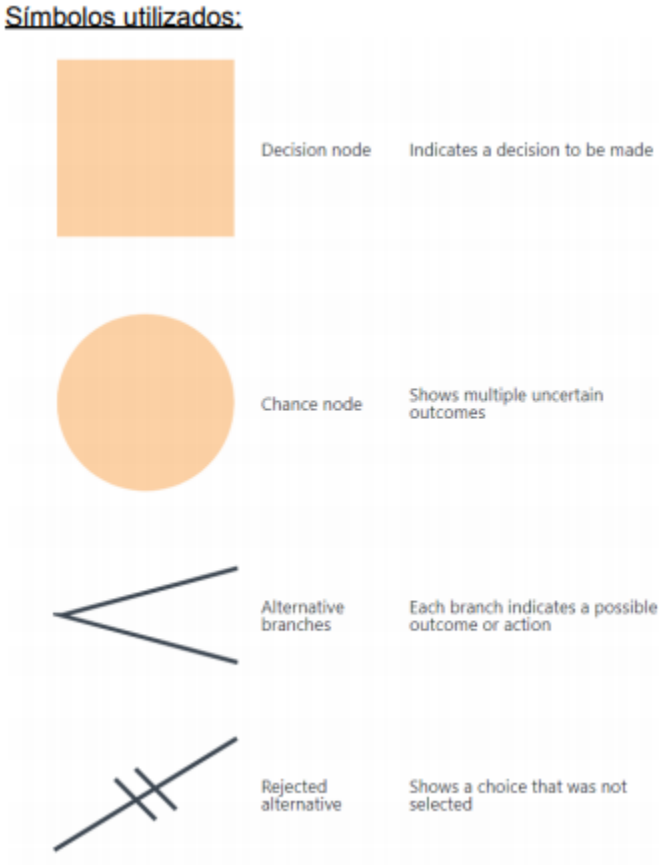
Descripción

Un árbol de decisión es un modelo para la predicción basado en diagramas de construcciones lógicas. funciona en base a declaraciones de control condicionales.

Imagen. 1. Árbol de decisión sencillo:



Imagenes 2. Símbolos utilizados en un decision tree



Funcionamiento

Un árbol de decisión funciona de manera similar a un diagrama de flujo, de modo que lo que sería un nodo en una diagrama de flujo en un árbol de decisión es una prueba donde las ramas que salen de este representan el resultado de esta prueba.

Un árbol de decisión tiene tres tipos de nodos:

1. Nodos de decisión: tienen dos opciones entre las cuales elegir.
2. Nodos de oportunidad: son nodos de decisión pero que pueden tener más de dosa posibles resultados.
3. Nodos terminales: indica el resultado final del arbol de decisión.

Otro aspecto importante de los arboles de decisión son las reglas de decisión, en este el resultado forma parte del nodo y la cláusula esta conformada por las ramas y el nodo, de modo que if condicion1 and condicion2 and condicion3 then outcome. Las reglas de decisión se pueden
generar de las reglas de asociación.

Variable Respuesta:

Con la respuesta seleccionada podemos tener bastante información y ver los factores que determinaron la muerte de una persona en su edad (Variable de respuesta: Edad). Por ejemplo, podemos tener un modelo que nos indique en cuanto a qué factores importan más para saber si una persona de tercera edad fallece o cuales son los factores para los niños menores de un año que aparecieron en nuestro dataset y saber que cosas hay que mejorar qué características tienen y cómo lo podemos combatir. Para explicar un poco mejor la variable de respuesta se necesita entender cómo funciona, este permite minimizar cualquier tipo de criterio de error, por ello se desea minimizar la suma de cuadrados residuales, ya que es una medida de error utilizada en la configuración de regresión lineal, pero para ello se requiere una diversa tecnología computacional, ahí se debe de considerar todas las particiones posibles del espacio de la característica, por lo tanto se debe utilizar un enfoque de búsqueda menos intensivo en computación, es más sofisticado y aquí es donde se utiliza el regression tree.

El algoritmo comienza en la parte superior del árbol y dividiendo el árbol en dos ramas únicas, lo que crea una partición de dos espacios. Lleva a cabo esta decisión en particular en la parte superior del árbol varias veces y elige la decisión de las características que minimiza la suma de cuadrados residuales, estas son particiones secuenciales del conjunto de datos que se van realizando para maximizar las diferencias de las variables dependientes, esto nos quiere decir que fue muy buena opción para calcular la tasa de mortalidad que se vive dia a dia.

Transformaciones del dataset y Obtención de conjuntos de entrenamiento y prueba:

Con el objetivo de tener solo información relevante y procesable es necesario trabajar un poco toda la información que poseemos. La base de datos con la que se trabajó fue la de defunciones que ofrece el INE en el rango de años 2009-2020. Lo primero para llevar a cabo fue separar las variables que nos son útiles de las que no. En este proceso se descartaron entonces las siguientes variables: Mupreg (Municipio de registro), mesreg (Mes de registro), Añoreg (Año de registro) Depocu (Departamento ocurrencia), Mupocu (Municipio ocurrencia), Areag, caudef.descrip (Descripción de la causa) year (Año), Diaocu (Día de ocurrencia), Añoocu (Año de ocurrencia), Nacdif (Nacimiento del difunto), Ciuodif (Ciudad del difunto) estas no fueron halladas útiles, pues, más sólos son detalles del fallecido y no complementan con la causa o la edad.

Siendo las siguientes las columnas con las que se trabajó:

Imagen 3. Columnas utilizadas en el proyecto.

```
defunciones.columns
```

```
Out[7]: Index(['Depreg', 'Sexo', 'Mesocu', 'Edadif', 'Getdif', 'Ocudif', 'Mredif',
       'Caudef', 'Asist', 'Ocur', 'Escodif', 'Pnadif', 'Predif'],
```

Ahora bien, para tener un mejor control y poder tener una mejor aproximación a lo que se extrae de los datos, se decidió dividir en 6 grupos lo que viene siendo nuestra variable respuesta y crearon etiquetas para cada uno, quedando así:

Imagen 4. Etiquetas de variable respuesta.

```
data['AgeRange']= 0
data.loc[data.Edadif == 0, 'AgeRange'] = 0
data.loc[(data.Edadif > 0) & (data.Edadif < 11), 'AgeRange'] = 1
data.loc[(data.Edadif >=11) & (data.Edadif < 18), 'AgeRange'] = 2
data.loc[(data.Edadif >=18) & (data.Edadif < 28), 'AgeRange'] = 3
data.loc[(data.Edadif >=28) & (data.Edadif < 45), 'AgeRange'] = 4
data.loc[(data.Edadif >=45) & (data.Edadif < 71), 'AgeRange'] = 5
data.loc[data.Edadif >=71, 'AgeRange'] = 6
```

```
dictionaries_age = {
    '0': 'Recien nacidos',
    '1': 'Niños',
    '2': 'Adolecentes',
    '3': 'Jovenes',
    '4': 'Adulto',
    '5': 'Adulto mayor',
    '6': 'Anciano'
}
```

Adicional a esto, también se eliminaron filas donde hubiesen datos vacíos. Igualmente, para que la información acerca de las causas de enfermedades fuese más fácil de interpretar , se cargó un diccionario que reemplazó los códigos durante el trabajo. Tras haber obtenido las 10 primeras enfermedades gracias a un gráfico de barras en el que se pudo determinar que después de la décima enfermedad, los casos descendían por debajo de los 1500 casos quedándose cortos en comparación con las primeras 10, luego de haber obtenido la lista de las enfermedades más presentes en el dataset, se mapearon los valores de manera que tuvieran un valor numérico que los representase para poder trabajar más fácil con estas.

Se procedió a sacar las 10 ocupaciones más presentes en el dataset, al igual que con las enfermedades estás venían sólo identificadas con un código por lo que se volvió a usar un diccionario para entender sus valores durante el proceso de desarrollo, tras ver que después de estas 10 el resto de ocupaciones es muy pequeña en proporción a estas, así que tras volver a

mapear los datos a un valor numérico que lo represente se filtraron solo las ocupaciones más comunes para tener datos más representativos.

Por último se pudo apreciar que habían datos inexactos en el dataset, en el caso de las filas que tuvieran un valor 0, se decidió llenar esto dado que de ser una variable categórica no se tomaría en cuenta en un conteo, por otro lado el valor 9999 que era un valor recurrente en el dataset para denotar la falta de información se limpió completamente.

Algoritmos seleccionados y resultados:

Tanto como el el random forest y en el decision tree armamos árboles de decisión que tanto son se utilizan para problemas de clasificación de regresión como de clasificación y estos nada más son un árbol en base a decisiones y resultados para llegar a un outcome específico. Este algoritmo nos otorga una manera visual simple para poder ver los factores de mayor importancia en nuestro dataset y saber bien de qué depende nuestra variable de respuesta que al final lo que queremos saber con esta es qué factores determinan la muerte de un individuo en su Edad.

La razón por la cual elegimos los dos es que uno es mucho más pesado de calcular que otro en el random tomamos el promedio de las decisiones tres y ya se arma el modelo a partir de eso. Queremos medir que tan importante sería en este caso y comparar su performance con un dataset bastante largo y si realmente valió la pena comparando las matrices de confusión, también encontramos un paper donde utilizaron una mezcla entre linear regression y decisión tree para poder medir la edad en las muertes y hablaban de cómo los árboles de decisión podrían manejar el approach en caso hayan factores no lineales en su predicción e interacciones en una dimensión superior que no se logra notar a partir del modelo con regresión lineal y según lo que vimos en nuestro dataset tenemos mucha información que no se comportaría de manera lineal así como el departamento de la persona o si tuvo asistencia médica o no lo cual nos permitirá una buena predicción.

Además, a nivel de implementación tienen también sus beneficios, por ejemplo en el caso del regression tree, se puede dar prioridad a un criterio por encima de otros en caso que no todas las variables tengan un mismo peso, así mismo, este tampoco requiere por ejemplo, que se normalice la información para trabajar con ella, encima de esto, y para fines de este proyecto, es más fácil extraer la información y graficarla de manera entendible incluso para aquellos que no estén versados en el tema.

Discusión de los resultados :

Discusión de la matriz de confusión árbol de decisión

Accuracy del modelo

Nuestro modelo presentó una presición del 63.62% no siendo excelente pero siendo decente para el modelo que hemos elegido. Aunque definitivamente es algo que se puede mejorar.

Árbol generado

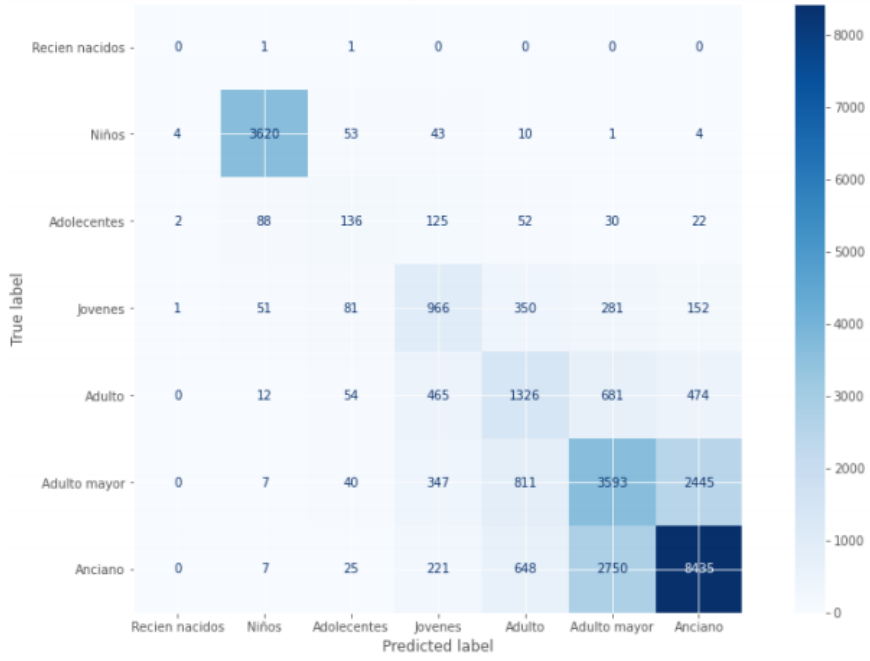
El árbol nos muestra que el primer valor que determina la predicción de un fallecido es la causa de defunción así que podemos asumir que hay una fuerte relación entre la edad y la causa de enfermedad entre las 10 más comunes. Esto nos puede servir para determinar qué es un factor importante a tomar en cuenta cuando alguna persona con mezcla (edad + enfermedad que presente riesgo) a esa edad sea evaluada y también esta información se puede empezar a tomar gracias a esto. Además observamos que la ramificación es cuando la enfermedad pertenece a las primeras dos y la otra ramificación es las restantes. Hay dos enfermedades que determinan bastante dentro de nuestro árbol. Nuestro árbol decidimos asignarle una profundidad máxima en sus ramas de 5 debido a que era muy pesado plotear el árbol completo pero sentimos que esta profundidad nos permitirá tener como mínimo 5 valores realmente importantes dentro de la relación de edad - defunción. Dentro de nuestro árbol de decisión en el primer rango es tan grande que el árbol decide volver a ramificar en base de la causa esto siempre dentro de la ramificación de la izquierda. Y luego ya pasamos a otros valores diferentes a la causa que serían la ocurrencia y el departamento nos llama la atención que el departamento sea un factor importante para la predicción de la edad esto puede ser debido a una de las enfermedades dentro de esta ramificación es Neumonía y como habíamos visto en nuestros clusters había una relación fuerte de esta enfermedad con los departamentos frios ademas que el sitio de ocurrencia es un factor prácticamente separando a cuando es en una casa de salud o un domicilio lo cual tiene sentido. Dentro de estas causas incluyendo ahora la diabetes observamos que el sexo es un factor para determinar la edad de defuncion. Por otro lado, observamos que nuestro árbol nos indica que el mes de ocurrencia puede ser un factor seguido de la escolaridad y la etnia del difunto. Esto nos hace pensar que probablemente en el siguiente nivel podriamos observar la ocupacion justo debajo de la escolaridad o incluso debajo de la etnia al ser concientes que en Guatemala existe un problema de discriminación. También nos fijamos que hay enfermedades

Matriz de confusión

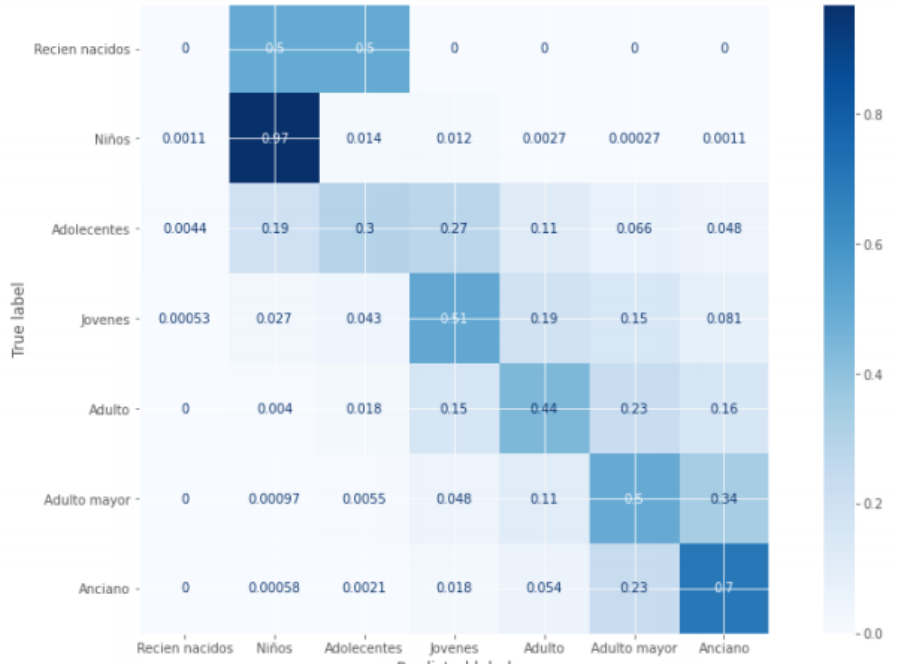
Dentro de nuestra matriz de confusión normalizada podemos observar que nuestras predicciones de los niños y ancianos fueron los más acertados con 97 y 70 % de acierto

respectivamente. Por el otro lado observamos que los adolescentes tuvieron el más bajo promedio de aciertos siendo confundidos en su mayoría con niños y con jóvenes esto pudo haber pasado por dos razones nuestro rango no está bien establecido o la muestra no es lo suficientemente confiable para poder predecir esto. Para corregirlo podemos estratificar los datos asegurándonos que todos tengan de cada uno de los grupos un porcentaje. El resto de los grupos que obtuvimos observamos en promedio que alrededor de un 50% de acierto se
tuvo. Como dato importante podemos notar que los recién nacidos tuvieron bastante confusión con los adultos mayores consideramos que esto puede ser al delimitar el rango en un valor únicamente (0 años) o bien puede ser que los valores en este dataset no muestran tener una relación realmente con este grupo y no haya factores para predecir la muerte de un recién
Nacido.

Matriz no. 1. Matriz de confusión no normalizada, árbol de decisión.



Matriz no. 2. Matriz de confusión normalizada, árbol de decisión.



Discusión de la matriz de confusión y resultados con random forest

Accuracy del modelo

Lo primero que podremos observar es una mejora en nuestro accuracy de un 4%. No fue algo realmente significativo para tener una muestra de 1000 árboles dentro de nuestro random forest de clasificación. El modelo presentó un 67.17% de acierto.

Árbol generado

Este modelo no podemos mostrar los árboles ya que son 1000 y es en base al promedio.

Mejoras

Donde observamos mejoras es en la cantidad de niños acertados miramos un incremento del 96% -> 98% la diferencia de los resultados en esto provienen en que ya no hubo confusión con Recién nacidos y Adolescentes.También los ancianos 70% -> 79% incrementaron un 9%. Esta mejora es dada en que hubo menor confusión en un 5% con los adultos mayores.

Empeoramientos

Los adolescentes sí se vieron afectados de un 30% a un 26% La predicción de los Jóvenes, Adultos, y adultos mayores vieron una pérdida en su predicción no muy significativa la mayoría de un 1 % pero llama la atención.

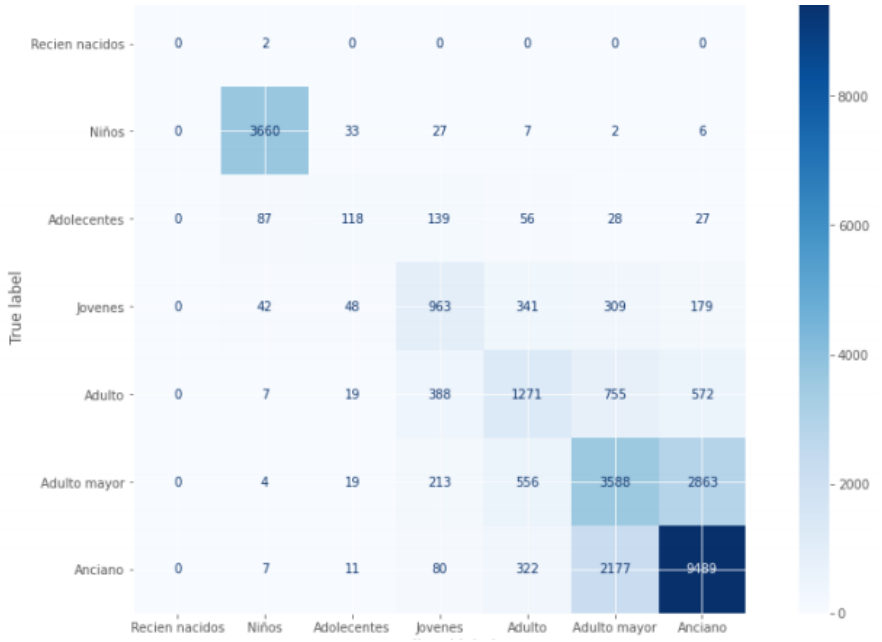
Diferencias

Podemos observar que con los adultos mayores se tuvo una menor confusión con los adultos en general pero más con los ancianos. También en el grupo de jóvenes nuestro modelo logró Discernir mejor entre adolescentes y jóvenes pero empeoro con los adultos. Esto podría ser debido a overfitting y generalización.

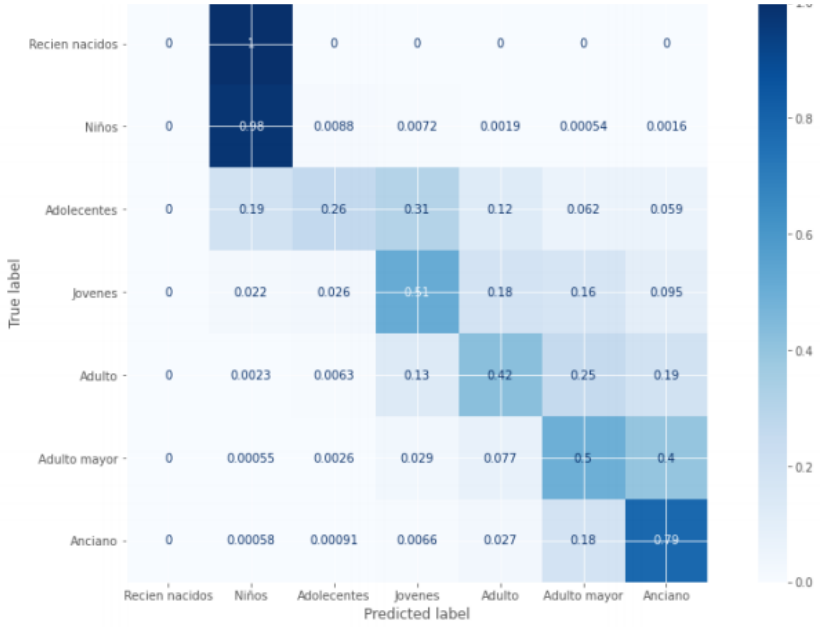
Resultados aislados

Viendo nuestra matriz miramos justamente que la concentración de datos está en los niños, los adultos mayores y los ancianos. Por ejemplo los recién nacidos apenas hay 3 dentro de la muestra y ni uno fue bien predecido

Matriz no. 3. Matriz de confusión no normalizada, random forest.



Matriz no. 4. Matriz de confusión normalizada, random forest.



Arbol de decisión con data estratificada

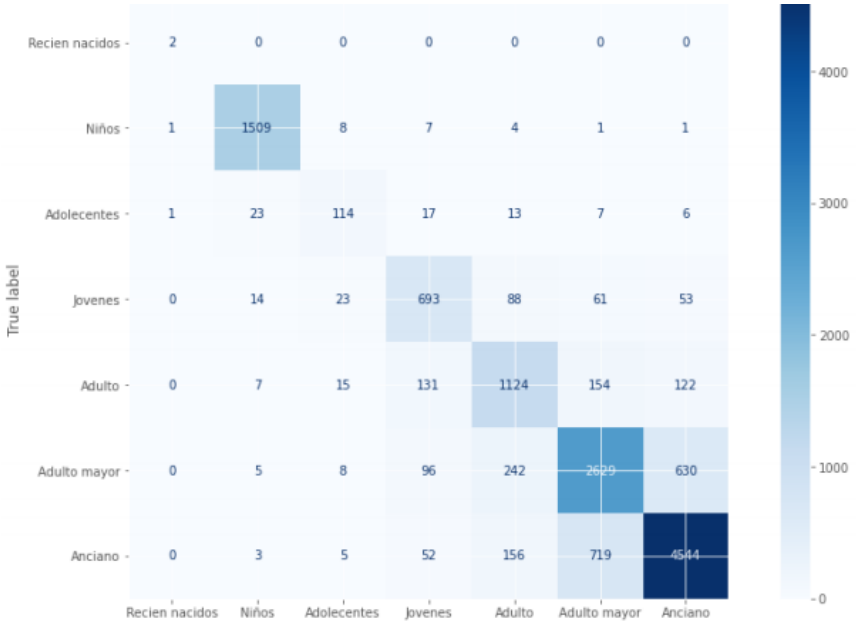
Accuracy del modelo

Nuestro modelo en precisión decayó un poco, podemos observar que bajó casi un 10% y viéndolo únicamente desde este punto la estratificación no fue de gran ayuda. Pero falta examinar
la matriz de confusión para ver realmente que paso

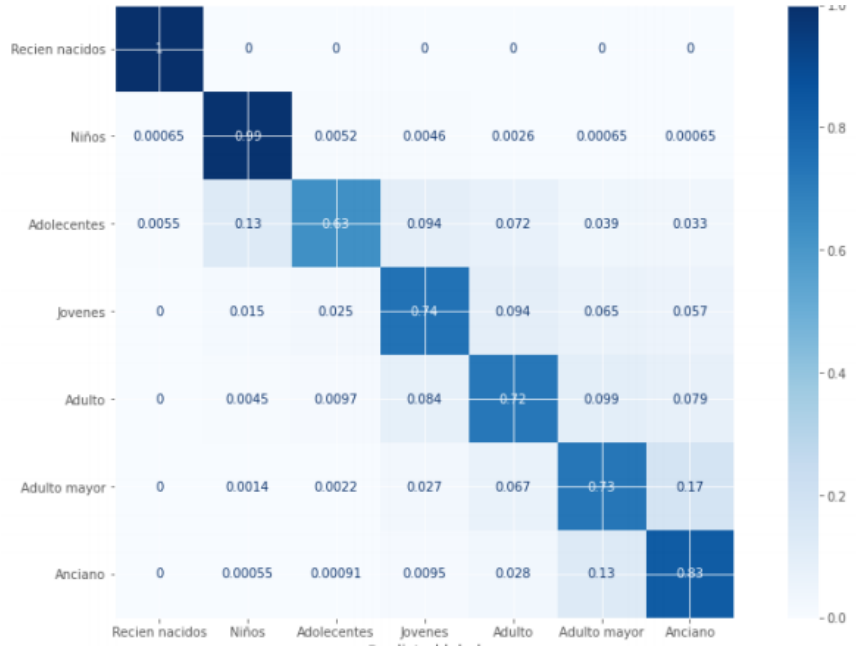
Matriz de confusión

Cuando visualizamos nuestra matriz de confusión podemos observar que efectivamente hemos tenido prácticamente una mejora en la predicción en todos nuestros labels. Creemos que
nuestro accuracy puede verse afectado por que la distribución de la muestra cambio y en el modelo sin estratificar casi todos eran ancianos y ahí hubo un buen % de acierto. Dos modelos
que podemos observar a simple vista que mejoraron de gran manera fue los recién nacidos con un 100% de éxito aunque fue una muestra unicamente de 2 niños, el otro seria los
adolescentes mejorando de un 30% a un 63% reduciendo en gran parte su confusión con los niños y jóvenes.

Matriz no. 5. Matriz de confusión no normalizada, árbol de decisión (data estratificada).



Matriz no. 6. Matriz de confusión normalizada, árbol de decisión (data estratificada).



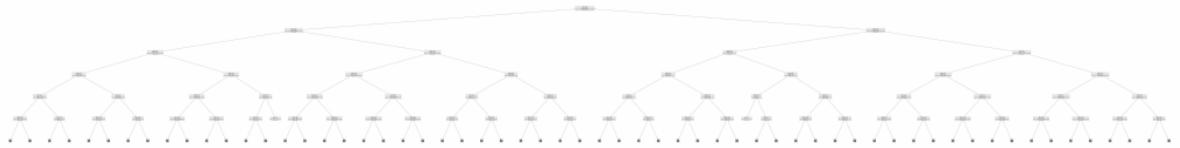
Mejoras

Este modelo mejoró en todos los aspectos posibles dentro de la matriz de confusión. Es un buen ejemplo porque no solo nos tenemos que fiar del accuracy para saber si un modelo es
Bueno si no una matriz de confusión nos indicará de mejor manera como se comporta el modelo con todos los casos en cuanto a la clasificación. Esto es para algoritmos de clasificación.

Análisis del árbol generado

Dentro del árbol notamos que la primera variable ahora es la ocupación del difunto. A diferencia del árbol no estratificado donde nos mostraba que el primer valor a tomarse en cuenta era la causa acá lo notamos también pero con un poco menos de relevancia aunque sigue siendo relevante aparece en el tercer nivel de la ramificación izquierda. Lo primero que notamos es que la primera división se hace justo con peones de explotación y oficios domésticos no remunerados. De este grupo notamos que la edad en que muere según nuestro modelo depende del sitio de ocurrencia específicamente si fue en un centro de salud u hospital o si fue en otra ubicación empezamos a notar que acá hay la causa empieza a ser factor de si fue un infarto, diabetes o neumonía no especificada. Cuando ya es una neumonia vemos que dependen de si fue en un hospital o nada mas en un centro de salud y en un infarto depende de si fue en un hospital publico o privado o sie fue diabetes empieza ya a pesar factores como el departamento y el sexo. Del otro lado dentro de esta ramificacion notamos que factores como si fueron lesiones o muertes sin asistencia empiezan a relacionarse con el sexo. Es interesante ver cómo es un factor tan importante en especial para estos trabajos de explotación. Del otro lado notamos que cuando son infantes su escolaridad depende como factor de la edad lo cual tiene sentido hay una relación fuerte en el grado y la edad de un difunto pero a partir de esto empezamos a notar que de nuevo los infartos y neumonía empiezan a ser factores como el mes de ocurrencia como en meses más fríos esto puede afectar. De ahi notamos que sifue en un hospital o centro medica dependiendo de si fue de nuevo un ataque, diabetes o neumonia el sexo empieza a ser un factor mientras que si fue otra causa de muerte ya tenemos el grupo delos agricultores. En general notamos que quienes mandan acá en estos niveles son el sitio + la ocupación + la causa de la defunción.

Imagen 5. Árbol de decisión data estratificada



Cual fue mejor?

Random Forest vs Árbol de Clasificación

Creemos nosotros que para tener un resultado más acertado definitivamente un random forest será lo que mejor nos irá pero hay que recordar que este es más caro y nos genera una colección de árboles.

Árbol de Clasificación (estratificado) vs Árbol de Clasificación (no estratificado)

Respecto entre los dos modelos de árbol de decisión generados preferimos el modelo con la data estratificada ya que su acierto está mejor distribuido según las matrices de confusión. La

estratificación nos permite entrenar al modelo de una manera proporcionalmente igual y aprendiendo de todas las posibles variables de respuesta además que es importante resaltar que en el caso de que no se estratifique además el modelo también podría crashear ya que si por ejemplo hay 10 bebés en el dataset y ni uno aparece en el train set luego con test set habrá una diferencia de niveles

Veredicto

Para tener los mejores resultados se debe de hacer un random forest con la información estratificada obtendremos mayor precisión para las diferentes clases en nuestra variable de respuesta. Ademas que nos permitirá árboles más acertados para identificar factores en la edad de un difunto. En cuanto a tiempo, realmente no hubo diferencia entre los árboles solo con el random forest. Dependerá probablemente no solo del tiempo si no que tanta diferencia hay entre el random forest para determinar si vale la pena.