Avance 1: Modelo de machine learning para la predicción de intentos de suicidio en la ciudad de Medellín

Estudiantes

Santiago Escobar Casas 1214746431

Santiago Otálvaro Ospina 1020492578

Tutor

Javier Fernando Botia Valderrama

2508205 - Modelos de sistemas

Universidad de Antioquia

Ingeniería de Sistemas

2020-1

Descripción de los avances

Este primer avance corresponde a la definición de entradas y salida para los modelos que se usarán en etapas posteriores y a la limpieza del dataset, del cual se obtendrán otros 3.El procedimiento para la limpieza de datos es eliminar aquellas columnas que presentan un alto número de valores nulos (ya que al ser tan altos no sería razonable imputar datos) y depurar aún más los datos reduciendo "drásticamente" el volumen de registros con respecto al archivo original, pero conservando características que se consideran relevantes para la predicción.

Limpieza de datos

Inicialmente se habia propuesto predecir la posibilidad de que una persona cualquiera intentara el suicidio basados en unas condiciones previas, pero para realizar esto necesitaríamos de datos de caracterización de personas que no han intentado el suicidio, el "dataset"(DS) con el que se está trabajando para el caso exacto de Medellín no presenta esta condición, solo presenta información de personas que ya han intentado suicidarse, por lo que debemos replantear un poco lo que se va a predecir(salida).

Eliminación de columnas no relevantes para el problema

Al realizar el análisis de los datos presentes en el DS se decidió tomar como salida la columna 'inten_prev', la cual hace referencia a si la persona ha intentado previamente el suicidio, salida será predicha a partir de las condiciones previas referentes a problemas o trastornos ['prob_parej', 'enfermedad_cronica', 'prob_econo', 'muerte_fam', 'esco_educ', 'prob_legal', 'suici_fm_a', 'maltr_fps', 'prob_labor', 'prob_consu', 'hist_famil', 'idea_suici', 'plan_suici', 'antec_tran', 'tran_depre', 'trast_personalidad', 'trast_bipolaridad', 'esquizofre', 'antec_v_a', 'abuso_alco']

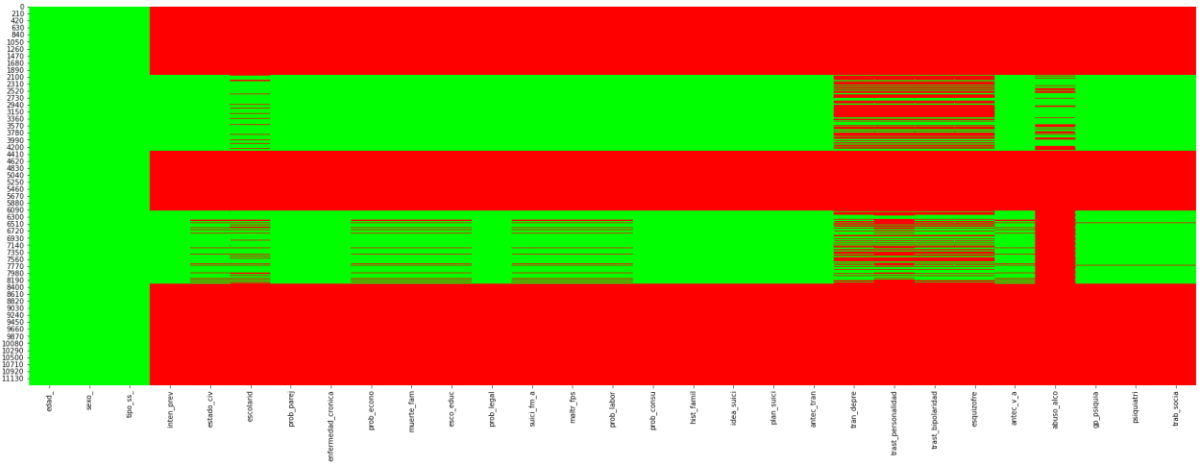
Con la decisión de definir como salida 'intent_prev' se pretende hallar aquellas a persona que se encuentran en lo que se podría definir como un "grupo de mayor riesgo", así mismo,extrapolar para hallar las situaciones o precondiciones más riesgosas, esto en aras de no alejarse mucho del problema propuesto inicialmente aun teniendo en cuenta las dificultades presentadas con respecto al DS , adicionalmente, se descubrió que una columna muy parecida 'intentos', la cual especificaba el número de intentos de suicidios previo de una persona, fue tenida en cuenta como posible salida pero finalmente se descartó, ya que aproximadamente el 85% eran valores nulos

Para este caso no es relevante la forma de intento de suicidio por lo que se eliminaran las columnas ['ahorcamien','arma_corto','arma_fuego','inmolacion','lanz_vacio','lanz_vehic','lanz_agua','intoxicaci'], tampoco son de interes para el problemano las columnas ['id','cod_ase_','tip_cas_']. Adicionalmente ['uni_med_', 'evento'] se eliminan ya que, en la primera solo 1 registro tiene un valor diferente y en el caso de la segunda 'evento' solo hay un valor posible. Finalmente la columna ['intentos'] presenta un alto porcentaje de valores nulos por lo que también se eliminará.

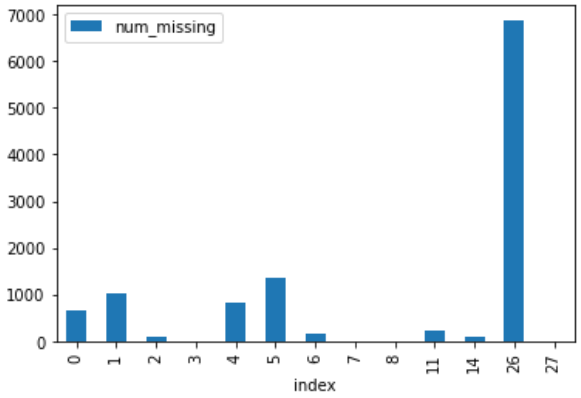
Para este caso no es relevante la forma de intento de suicidio por lo que se eliminaran las columnas ['ahorcamien','arma_corto','arma_fuego','inmolacion','lanz_vacio','lanz_vehic','lanz_agua','intoxicaci'], tampoco son de interes para el problema las columnas ['id','cod_ase','tip_cas', 'pac_hos', 'nombre_barrio', 'comuna'], ni se tiene especial interes en las fechas por lo que se eliminaran ['semana', 'fec_con', 'ini_sin', 'fec_ocurr', 'year'].Finalmente ['uni_med', 'evento', 'intentos'] se eliminan ya que, en la primera solo 1 registro tiene un valor diferente ,en el caso de la segunda solo hay un valor posible y la tercer presenta un alto porcentaje de valores nulos.
Las columnas que quedan son aquellas que inicialmente se pretendían usar como entrada y algunas adicionales las cuales se explorarán para descubrir si existe alguna correlación o si son relevantes para la salida definida.

Datos faltantes

El siguiente paso es el borrado de datos faltantes. Por medio de un mapa de calor podemos observar, de manera gráfica cuales son los intervalos de información que faltan(datos nulos en rojo) en el DS.



En la siguiente gráfica se muestra la cantidad de datos nulos por fila o registro (eje X cantidad de datos nulos y eje Y número de filas).

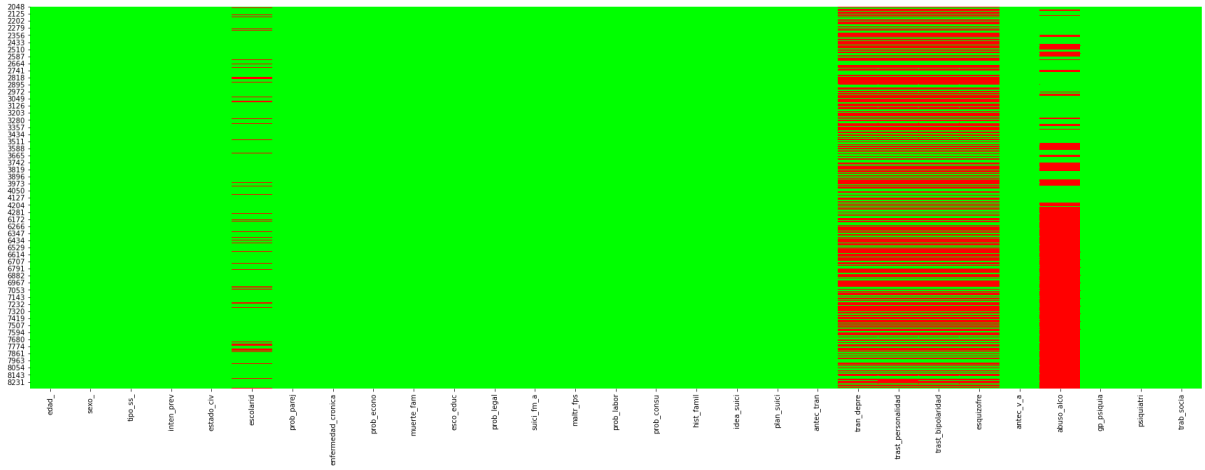


Como se puede observar, hay aproximadamente 7000 filas con 26 datos nulos. La solución que se plantea para el problema con los datos faltantes es:

1. Eliminar registros con 11 o más datos faltantes

2. Eliminar registros con información faltante en 'inten_prev' (salida definida)
3. Tratar columnas restante, ya sea eliminandolas(aquellas con alto porcentaje de nulos) o imputando datos

Posterior a la barrido de datos nulos(paso 1), se obtiene la siguiente gráfica de datos faltantes con la respectiva tabla de porcentajes de datos nulos.



escolarid	9.93%	trast_bipolaridad	57.12%	gp_psiquia	0.14%
tran_depre	57.12%	esquizofre	57.12%	psiquiatri	0.19%
trast_personalidad	57.38%	abuso_alco	60.23%	trab_socia	0.19%

En este punto, se contemplan tres opciones:

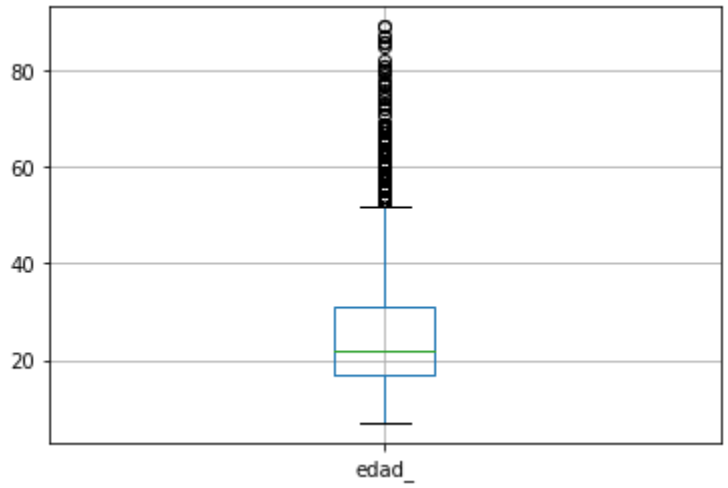
1. Eliminar aquellas columnas que presentan alto número de valores nulos [tran_depre, trast_personalidad, trast_bipolaridad, esquizofre, abuso_alco], ya que al ser tan altos no sería razonable imputar datos.
2. Depurar aún más los datos y reducir "drásticamente" el volumen de registros con respecto al original pero conservando algunas características que se considera podrían llegar a ser relevantes para la predicción.
3. Imputar dichos valores nulos, aun cuando son mayoría.

Se obtendrán tres datasets (cds, cds_few, cds_imputed), cada uno resultado de aplicar respectivamente una opción de las comentadas previamente, con el fin de compararlos en siguientes etapas al momento de desarrollar los modelos.

Primero se procede a detectar outliers, valores no válidos y terminar de etiquetar y limpiar los datos para que lo último a realizar sea separarlo según se comentó anteriormente.

Detección y eliminación de outliers y valores no válidos

El único dato realmente numérico es edad, en su diagrama de caja se observa que hay algunos valores muy alejados del promedio, pero teniendo en cuenta el contexto son considerados como válidos.



Para las demás columnas(categóricas) se realiza la gráfica de barras para conocer su distribución y si existe algún dato no válido; tal es el caso para 'escolaridad', en el cual se presentan algunos registros con valor en 6(no definido en el DS), casos para los cuales se elimina el registro.

Codificación de los datos

Inicialmente los valores posibles que tienen las columnas son:

- 'edad_' : numérico
- 'sexo_' : M, F
- 'tipo_ss_' : C= Contributivo, S=Subsidiado, P=Excepción, E=Especial, N= No asegurado, I= Indeterminado/Pendiente
- 'estado_civ' : 1= Soltero(a), 2=Casado(a), 3= Unión libre, 4= Viudo(a), 5= Divorciado(a)
- 'escolarid' : 1= Preescolar, 2= Básica primaria, 3= Básica secundaria, 5= Media técnica, 7= Técnica profesional, 8= Tecnológica o técnica, 9= Profesional, 10= Especialización, 11= Maestría, 12= Doctorado, 13= Ninguno, 14= Sin información
- ['inten_prev', 'prob_parej', 'enfermedad_cronica', 'prob_econo', 'muerte_fam', 'esco_educ', 'prob_legal', 'suici_fm_a', 'maltr_fps', 'prob_labor', 'prob_consu', 'hist_famil', 'idea_suici', 'plan_suici', 'antec_tran', 'tran_depre', 'trast_personalidad', 'trast_bipolaridad', 'esquizofre', 'antec_v_a', 'abuso_alco', 'gp_psiquia', 'psiquiatri', 'trab_socia'] : 1= Si, 2= No

La mayoría de columna son categóricas, sin embargo el dataset obtenido ya las tenía codificadas, por lo tanto se usará así, pero donde sea necesario se cambiarán algunas etiquetas y se codificaran aquellas que lo necesiten.

- 'sexo_'-: binary encoding M:0 F:1
- 'tipo_ss_' : Se utilizara one hot encoding

Las demás columnas se usarán tal y como están.

Imputación

A las columnas ['gp_psiquia', 'psiquiatri', 'trab_socia'] se les realiza una imputación simple de datos usando la mediana, obteniendo como resultado que solo las columnas [‘escolarid’, ‘tran_depre’, ‘trast_personalidad ‘, ‘trast_bipolaridad’, ‘esquizofre’, ‘abuso_alco’] quedan con valores nulos, de las cuales escolaridad tiene aproximadamente 10% y las otras son mayores a 55% de datos nulos

Obtención de cds_imputed e imputación para escolaridad

Inicialmente se realizó una imputación multivariante(se tienen en cuenta las otras columnas para imputar), pero se abandonó este enfoque ya que no se logró imputar con técnica de mayor frecuencia(moda), por lo que los valores imputados podrían resultar siendo decimales, lo cual no tendría sentido al ser datos categóricos. En cambio se realizó una imputación simple usando estrategia de mayor frecuencia, generando así un DS que se llamará “cds_imputed” y del cual se tomará columna ‘escolarid’ (ya con los datos imputados) para generar los otros DS.
Posteriorme, con el DS anterior a la imputación de todos los datos excepto escolaridad, se procede a obtener dos DS nuevos:
Para el primero (“cds”) tomando los datos así como se encuentran pero borrando las columnas con un alto porcentaje de valores nulos(cds).
Para el segundo (“cds_few”) se eliminan los registros con datos faltantes, quedando solamente 655 registros.
Finalmente se exportan como csv los DS generados para que sean usados en las etapas siguientes.

Con este proceso de limpieza de datos se pueden orientar de mejor manera las cosas para utilizar la información en el modelo.

Código usado en el procedimiento disponible aquí.