Tecnológico Nacional de México

Instituto Tecnológico De Hermosillo

proyecto final

Arvayo Vasquez Jose Angel

Maestro: Eduardo Antonio Hinojosa Palafox

Grupo: S9A

Hermosillo, Sonora 14 de diciembre de 2022

Índice

Introducción

El abuso sexual infantil es un tipo de maltrato que ha sido considerado como uno de los problemas de salud pública más graves que deben afrontar las niñas, niños y adolescentes de las sociedades. Aunque, aproximadamente, uno de cada cinco menores en el mundo sufre algún tipo de abuso sexual, aún es un tema tabú sobre el que no se habla lo suficiente.

En la mayoría de los casos, el abuso sexual infantil supone una experiencia traumática para el que la padece, interfiriendo en su adecuado desarrollo y repercutiendo negativamente tanto en su estado físico como psicológico. Las consecuencias derivadas de tales actos no solo afectan a las víctimas y a sus familias, sino que también acaban repercutiendo a toda la sociedad en su conjunto.

La minería de datos es una herramienta que nos permite generar conocimiento dándole sentido a una gran cantidad de datos, encontrando patrones y anomalías en el comportamiento en los datos registrados para predecir resultados. Esto puede representar una gran ayuda en el campo psicosocial, encontrando patrones de comportamiento con los cuales podemos crear estrategias de prevención, en este caso el abuso sexual infantil.

Desarrollo

Problema a desarrollar

el abuso sexual infantil es una de las mayores preocupaciones de una sociedad. Para proteger a los infantes del abuso debemos identificar los factores potenciales que propician estas situaciones. Usamos este conjunto de datos a partir de estos pensamientos, contiene 8 preguntas que fueron tomadas en la India sobre el conocimiento que tiene la sociedad de medidas que se pueden tomar para la prevención del abuso infantil. Al ser un dataset previamente etiquetado utilizaremos un modelo de clasificación para datos etiquetados.

Descripción del conjunto de datos

El conjunto de datos consta de 8 columnas que consta de 7 preguntas con 2 posibles respuestas y un target que nos dice que tan amplio es el conocimiento de la persona, [“agree”, “disagree”] o [“yes”,”no”] según la pregunta, las 4 primeras hacen referencia a que tan enterada está la persona sobre el contexto social actual de la problemática, las siguientes 4 nos hablan:

1. "Children are safe among family members such as grandparents, uncles, aunts, cousins" [“agree”, “disagree”]
2. "Children are mainly abused by strangers in our society" [“agree”, “disagree”]
3. “Male children dont need sexual abuse prevention knowledge” [“agree”, “disagree”]
4. "Teaching sexual abuse prevention in school is not necessary. It will make children curious about sex" [“agree”, “disagree”]
5. Do you know what child grooming is? [“yes”,”no”]
6. Do you know what signs to look for to identify if your child has been abused? [“yes”,”no”]
7. Do you think children need post abuse counseling for recovering? [“yes”,”no”]
8. Do you think you should take legal action against the abuser of your child? [“yes”,”no”]
9. Knowledge Level

Programación y Modelado

1- importamos las librerías que vamos a utilizar, en este caso numpy nos ayuda a hacer cálculos con grandes cantidades de datos, pandas nos ayuda a importar, exportar y manipular sets de datos, seaborn nos ayuda a crear distintos tipos de gráficos, de sklearn importamos el modelo clasificatorio, la función de preprocesamiento, la función que nos permite separar el set de datos en datos de entrenamiento y evaluación, también las métricas de evaluación y la que nos permite crear la matriz de confusión.

```
[ ]  import numpy as np
     import pandas as pd
     import seaborn as sns
     from sklearn.preprocessing import LabelEncoder
     from sklearn.model_selection import train_test_split
     from sklearn.tree import DecisionTreeClassifier
     import matplotlib.pyplot as plt
     from sklearn import tree
     from sklearn.metrics import confusion_matrix
     from sklearn.metrics import classification_report
```

2-importamos el set de datos que describimos previamente

```
data=pd.read_csv('CSA-Data.csv')
```

3- Hacemos un análisis exploratorio de los datos, en el cual vemos que no hay valores nulos en todo el set, por lo tanto durante el preprocesamiento no tendremos la necesidad de hacer modificaciones al respecto.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3002 entries, 0 to 3001
Data columns (total 9 columns):
 #   Column                                                                                                Non-Null Count  Dtype
---  ------                                                                                                --------------  -----
 0   "Children are safe among family members such as grandparents, uncles, aunts, cousins"                3002 non-null   object
 1   "Children are mainly abused by strangers in our society"                                             3002 non-null   object
 2   Male children dont need sexual abuse prevention knowledge                                            3002 non-null   object
 3   "Teaching sexual abuse prevention in school is not necessary. It will make children curious about sex"  3002 non-null   object
 4   Do you know what child grooming is?                                                                  3002 non-null   object
 5   Do you know what signs to look for to identify if your child has been abused?                        3002 non-null   object
 6   Do you think children need post abuse counseling for recovering?                                     3002 non-null   object
 7   Do you think you should take legal action against the abuser of your child?                          3002 non-null   object
 8   Knowledge Level                                                                                      3002 non-null   object
dtypes: object(9)
memory usage: 211.2+ KB
```

4- Separamos el set de datos en variables y target, y al mismo tiempo creamos un vector con los nombres de las variables.

```
features = ['Children are safe among family members such as grand
X=data.iloc[:,:-1]
y=data.iloc[:,-1]
```

5- hacemos el preprocesamiento de los datos, en donde usaremos la función label encoder para transformar las variables datos numéricos con los cuales pueda hacer cálculos la computadora, para aplicarlo en todas las columnas usamos un ciclo for para automatizarlo

```
label = LabelEncoder()
for col in X.columns:
    X[col]=label.fit_transform(X[col])
X
```

```
<ipython-input-33-dfdb58f401c4>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

6- también aplicamos el mismo proceso transformamos los datos del target.

```
y=label.fit_transform(y)
```

7- dividimos las variables y los target en 2, los valores de entrenamiento y los de evaluación del modelo, 67% de los datos serán utilizados como valores de entrenamiento mientras que el 33% restante lo usamos para evaluar el sistema, y le pedimos que alterne las filas.

```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=44, shuffle =True)
```

8-Creamos el modelo clasificatorio, en este caso usaremos el árbol de decisión, la elección fue hecha en base a la naturaleza de los datos y las respuestas binarias que nos da, le damos como parámetros el criterio de entropía y una profundidad máxima de 10 para evitar el sobreajuste del modelo y lo entrenamos con los valores de entrenamiento que generamos anteriormente.

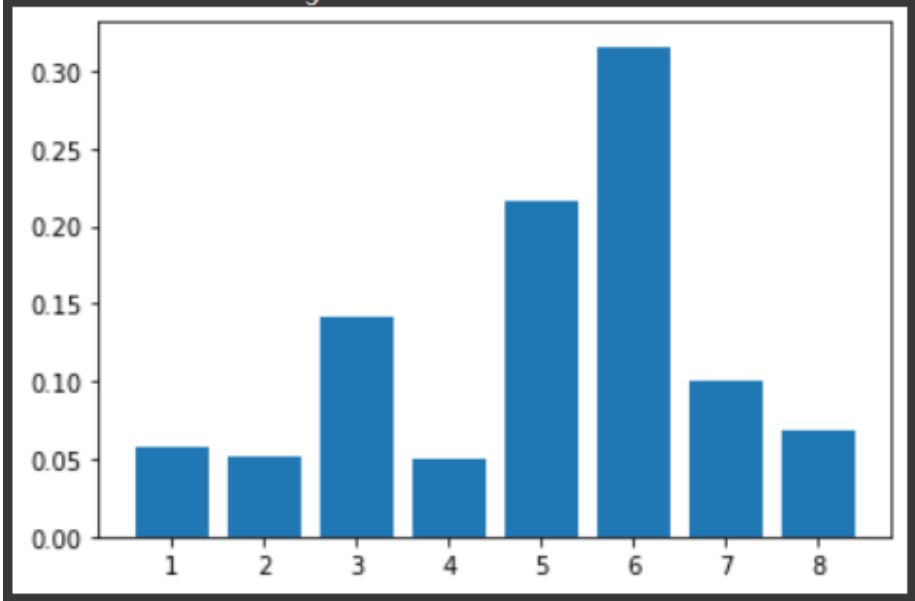
```
DecisionTreeClassifierModel = DecisionTreeClassifier(criterion='entropy',max_depth=10,random_state=33)
DecisionTreeClassifierModel.fit(X_train, y_train)
```

```
DecisionTreeClassifier(criterion='entropy', max_depth=10, random_state=33)
```

9-creamos un gráfico en el que vemos cuál es la prioridad que se les dio a las preguntas, en este caso la pregunta 6 (Do you know what signs to look for to identify if your child has been abused?) representa una gran importancia al momento de crear el árbol, de ahí sigue la pregunta 5 (Do you know what child grooming is?), luego la pregunta 3 (“Male children don't need sexual abuse prevention knowledge”), estas son las preguntas más significativas según el modelo para determinar el conocimiento de una persona sobre el abuso sexual infantil

```
X_bar=list(range(1,9))
plt.bar(X_bar,DecisionTreeClassifierModel.feature_importances_)
```

```
<BarContainer object of 8 artists>
```



10-generamos una predicción con los datos de testeo

```
y_pred_decisiontree = DecisionTreeClassifierModel.predict(X_test)
```

11- generamos el reporte con las métricas de rendimiento del modelo, el cual nos arrojó una precisión un poco mayor para las personas más informadas sobre el abuso infantil, en general tiene una precisión de .993 loq ue nos dice que es muy bueno

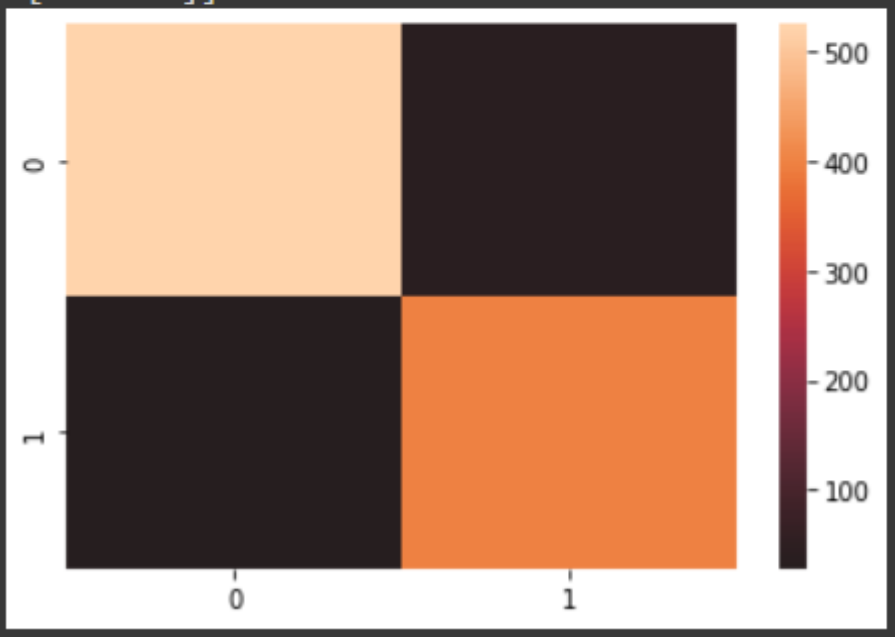
```
ClassificationReport_DT = classification_report(y_test,y_pred_decisiontree)
print('Classification Report : \n', ClassificationReport_DT )
```

```
Classification Report :
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       564
           1       0.91      0.93      0.92       427

    accuracy                           0.93       991
   macro avg       0.93      0.93      0.93       991
weighted avg       0.93      0.93      0.93       991
```

12- creamos una matriz de confusión para representar los datos que fueron clasificados de forma correcta e incorrecta. En esta ocasión nos encontramos que 525 personas fueron correctamente clasificadas como bajo conocimiento, 399 también fueron clasificadas correctamente como personas mejor informadas, pero por otro lado tenemos 39 personas fueron clasificadas como mejor informadas cuando en realidad no lo son, también 28 personas se clasificaron como peor informadas cuando en realidad sí están más informadas

```
SCM = confusion_matrix(y_test, y_pred_decisiontree)
print('Confusion Matrix : \n', SCM)
sns.heatmap(SCM, center = True)
plt.show()
```

```
Confusion Matrix :
 [[525  39]
 [ 28 399]]
```



13-importamos la librería para crear una imagen del árbol creado y posteriormente lo creamos, en los parámetros metemos el modelo clasificatorio y usamos los nombres de las columnas que generamos en un vector durante la separación de variables y target. Al ser demasiado grande incluimos una liga a la imagen generada

```
import graphviz
from sklearn.tree import export_graphviz


dot_data = export_graphviz(DecisionTreeClassifierModel, out_file=None,
                           feature_names=features,
                           class_names=True,
                           filled=True, rounded=True,
                           special_characters=True)
graph = graphviz.Source(dot_data)
```

file_name (1).png

Proyecto completo

https://github.com/JoseArvayoITH/proyecto-final-mineria

Resultados

Después de todo el proceso nos dio como resultado un modelo clasificatorio muy bueno para este caso, puede ser utilizado para hacer una encuesta masiva a lo largo del país y obtener información muy fidedigna sobre las áreas de oportunidad y deficiencias educativas para prevenir el abuso infantil, de esta manera se pueden tomar acciones más precisas y enfocadas para hacer campañas de concientización e información. Por otro lado nos abre la posibilidad de seguir expandiendo el análisis a modelos más complejos con más aristas donde se tomen en cuenta los contextos culturales, religiosos y económicos en los que viven las personas que contestaron la encuesta y cómo estos afectan el acceso a información que tienen.

Conclusiones

con base a lo que nos arrojó el modelo clasificatorio nos dice que una gran parte de la sociedad no está correctamente informada sobre las causas y formas de prevenir el abuso infantil, por lo tanto se hace la recomendación de hacer campañas de concientización sobre las causas del abuso infantil, del mismo modo se pueden implementar programas de educación sexual durante el desarrollo como forma preventiva al abuso, del mimmo modo implementar campañas de información sobre la importancia de la educación sexual ya que este es muy controversial en la sociedad debido principalmente a la desinformación.

Durante esta actividad final pudimos poner en práctica los conocimientos adquiridos durante todo el semestre, tanto la selección de los datos, el preprocesamiento de estos, determinar cuál es el modelo clasificatorio más adecuado para cada caso, crear el modelo y evaluarlo.

Referencias

"'Data mining', definición, ejemplos y aplicaciones - Iberdrola". Iberdrola. https://www.iberdrola.com/innovacion/data-mining-definicion-ejemplos-y-aplicaciones#:~:text=La%20minería%20de%20datos%20se,personalizadas%20de%20fidelización%20o%20captación. (accedido el 15 de diciembre de 2022).

"El abuso sexual infantil, un problema que afecta a toda la sociedad | Fundació Vicki Bernadet". Fundació Vicki Bernadet. https://www.fbernadet.org/es/el-abuso-sexual-infantil-un-problema-que-afecta-a-toda-la-sociedad/ (accedido el 15 de diciembre de 2022).