Segundo Informe Taller – Minería Datos

Presentado Por:

Yuri Natalia Bernal Mora

Nicolas Felipe Reyes Carillo

María Alejandra Maldonado Rojas

Modelo de predicción

Facultad De Ingeniería

Universidad Cooperativa De Colombia

Fredys Alberto Simanca Herrera

Abril de 2021

Contenido

1. Introducción

En la actualidad existe una gran variedad de trastornos mentales, cada uno de ellos con distintas manifestaciones. Los cuales normalmente se caracterizan por una combinación de alteraciones del pensamiento, la percepción, las emociones, la conducta incluyendo la violación de reglas-normas (grupales y sociales) y conductas inapropiadas para la edad (conducta sexualizada a los 7-8 años), incluyendo desobediencia a padres y profesores y alteraciones en todo tipo de relaciones.

Los más conocidos son la depresión, trastorno afectivo bipolar, esquizofrenia, psicosis, demencia, discapacidades intelectuales y los trastornos del desarrollo; como el autismo, TDAH, TOP Y TC, pero todos estos trastornos mencionados anteriormente aparecen normalmente desde la edad temprana y los trastornos mentales entre los niños son más comunes de lo que se espera; estos se pueden manifestar como cambios serios y pueden afectar directamente en la forma en que los niños aprenden, se comportan o manejan sus emociones, lo que causa angustia y problemas para pasar el día. Muchos niños ocasionalmente experimentan miedos y preocupaciones o muestran comportamientos perturbadores. Si los síntomas son graves y persistentes e interfieren con las actividades de la escuela, el hogar o juegos.

Los trastornos mentales infantiles afectan a muchos niños y familias. Los niños y niñas de todas las edades y orígenes étnicos / raciales que viven en todas las regiones de los Estados Unidos padecen trastornos mentales. Basado en el informe del National Research Council and Institute of Medicine y en Colombia se ha observado es que el número de personas de 0 a 19 años que consultan por trastornos mentales y del comportamiento es cada día mayor. De 2009 a 2017 se atendieron 2.128.573 niños, niñas y adolescentes con diagnósticos con código CIE 10: F00 a F99 (que agrupa los trastornos mentales y del comportamiento), con un promedio de 236.508 de personas atendidas por año, la tendencia es al aumento de casos cada año.

A pesar de que las cifras siguen en aumento en todo el mundo y que es un problema de salud pública. Al momento del diagnóstico aún se usan métodos muy rudimentarios; los cuales retrasan el proceso de diagnóstico o en otros casos ni siquiera se percatan de las dificultades por las que están pasando los niños y jóvenes y algo primordial para

tratar estos trastornos es la detección temprana de estos mismos y claro que existen diferentes tipos de tratamientos para cada uno de estos trastornos por ejemplo en trastornos de la conducta está la Terapia Cognitivo Conductual (TCC) ,La terapia multisistémica (MST) y Régimen asistencial fuera de casa.
Pero uno de los mayores problemas actualmente en cuanto a los tratamientos es la gran diversidad de síntomas que presenta una persona con estos trastornos y que conlleva a confusiones y mayor tiempo de diagnóstico.

Por esta razón nuestro Modelo de predicción estará basado para la detección de trastornos de conducta en niños y adolescentes de 1 a 17 años. El cual tiene como objetivo general desarrollar un entorno controlado mediante la inteligencia artificial la cuál facilitará la detección precoz de todos estos trastornos y reducir los tiempos de diagnóstico y generar un avance e impacto social enfocado en psicología mediante la aplicación de IA.
Hoy en día ya existen muchas herramientas tecnológicas que favorecen los procesos terapéuticos ya que estos sistemas se basan en el funcionamiento observado en el sistema cognitivo humano.
La idea de iniciar con esta investigación es de llegar a poder implementar un ambiente controlado para el manejo de problemas psicológicos los cuales han de ser tratados con una inteligencia artificial, de la cual podríamos decir que puede cumplir de una mejor manera ciertas funciones en cuanto a la mente. la Inteligencia Artificial se podría dotar de funciones analíticas y de un aprendizaje constante el cual nos permitiría obtener soluciones a problemas mucho más eficientes "Sin entrar en lo ético, humano y psicológico".

2. Metodología

2.1. Modelos de predicción

Los modelos de predicción en la minería de datos vienen de la inteligencia artificial, estas técnicas son básicamente algoritmos más complejos y modernos que se aplican sobre una cantidad de información y que nos permite obtener resultados sobre el análisis de la información.

"Un modelo predictivo es un conjunto de procesos ejercidos a través de técnicas computacionales de análisis de datos que ayudan a inferir la probabilidad de que ocurran determinadas situaciones previas a si consecución" *España, R. (2020, 3 febrero). Qué es un modelo predictivo y cómo se aplica al negocio. Agencia b12 España. https://agenciab12.com/noticia/que-es-modelo-predictivo-como-aplica-negocio*

En este proyecto no vamos a crear modelos de predicción, lo que vamos a hacer es a utilizar un modelo de predicción que se adecue a nuestras necesidades según el objetivo del proyecto, para poder hacer una elección efectiva de este modelo tenemos que conocer los existentes y sus características principales, es importante mencionar que al ser un modelo y no una metodología nos estable que hacer y no como hacerlo.

2.1.1. Agrupamiento o Clustering

El clustering consiste en la agrupación automática de datos. Al ser un aprendizaje no-supervisado, no hay una respuesta correcta. Esto hace que la evaluación de los grupos identificados sea un poco subjetiva.

Clustering es una técnica de Machine Learning que implica la agrupación de puntos de datos. Dado un conjunto de puntos de datos, podemos utilizar un algoritmo de agrupación para clasificar cada punto de datos en un clúster específico. En teoría, los puntos de datos que están en el mismo clúster deben tener propiedades y/o características similares, mientras que los puntos de datos en diferentes clústeres deben tener propiedades y/o características muy diferentes

Existen diferentes tipos de agrupamientos como lo son:

Agrupamiento K Means

K Means es probablemente el algoritmo de clustering más conocido, es fácil de entender e implementar en código.

Agrupamiento Mean Shift

La agrupación Mean Shift es un algoritmo basado en ventanas deslizantes que intenta encontrar áreas densas de puntos de datos. Es un algoritmo basado en el centroide, lo que significa que el objetivo es localizar los puntos centrales de cada clúster, lo que funciona actualizando a los candidatos para que los puntos centrales sean la media de los puntos dentro de la ventana deslizante.

DBSCAN

DBSCAN es un algoritmo de agrupamiento basado en la densidad similar a Mean Shift, pero con un par de ventajas notables. Las siglas DBSCAN significan agrupamiento espacial basado en densidad de aplicaciones con ruido.

Agrupamiento utilizando modelos de mezcla gaussiana

Los modelos de mezcla gaussiana (GMM, por sus siglas en ingles) nos dan más flexibilidad que los K Means. Con los GMM suponemos que los puntos de datos están distribuidos por Gauss, esto es una suposición menos restrictiva que decir que son circulares usando la media. De esta manera, tenemos dos parámetros para describir la forma de los clústeres, la media y la desviación estándar. Tomando un ejemplo en dos dimensiones, esto significa que los clústeres pueden tomar cualquier tipo de forma elíptica, ya que tenemos una desviación estándar tanto en la dirección X como en la Y. Por lo tanto, cada distribución gaussiana se asigna a un solo grupo.

Agrupamiento Jerárquico

Los algoritmos de agrupación jerárquica se dividen en dos categorías: de arriba hacia abajo o de abajo hacia arriba. Los algoritmos ascendentes tratan cada punto de datos como un solo clúster al principio y luego fusionan sucesivamente o aglomeran pares de clústeres hasta que todos los clústeres se hayan fusionado a un único clúster que contenga todos los puntos de datos. Por lo tanto, la agrupación jerárquica ascendente se denomina agrupación algomerativa jerárquica. Esta jerarquía de clúster se

representa como un árbol o dendograma. La raíz del árbol es el único clúster que recoge todas las muestras, siendo las hojas los clústeres con una sola muestra.

2.1.2. Naive Bayes

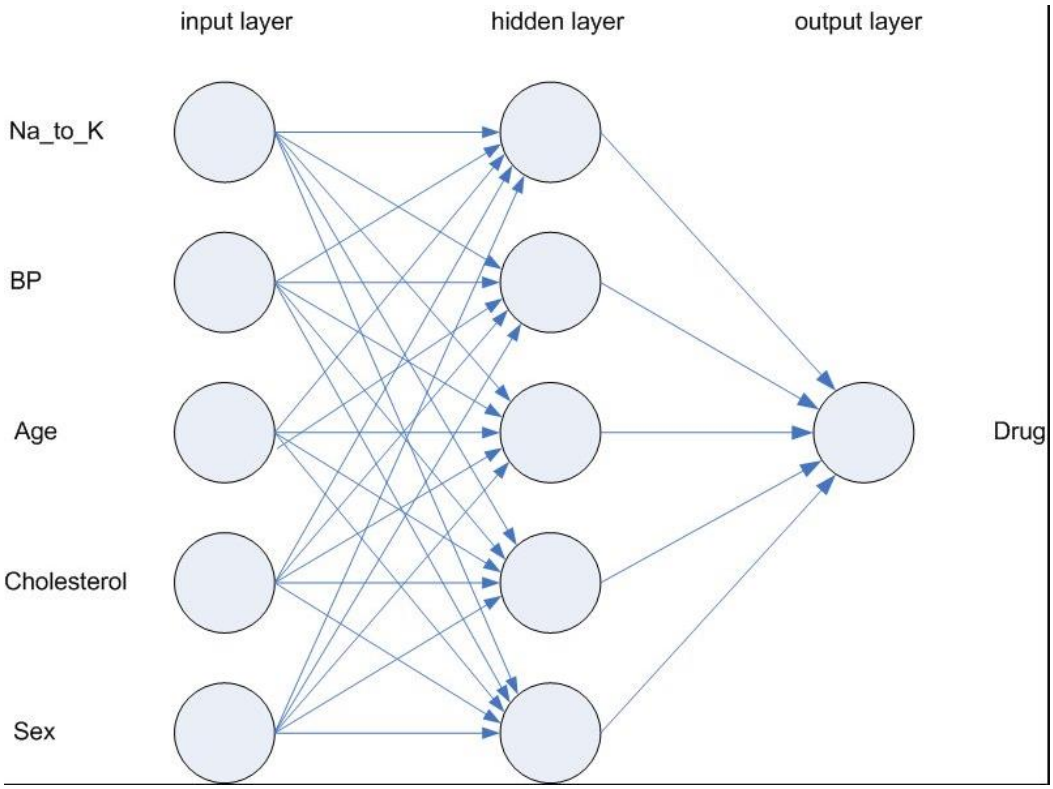
El algoritmo clasificador Naïve-Bayes (NBC), es un clasificador probabilístico simple con fuerte suposición de independencia. Aunque la suposición de la independencia de los atributos es generalmente una suposición pobre y se viola a menudo para los conjuntos de datos verdaderos. A menudo proporciona una mejor precisión de clasificación en conjuntos de datos en tiempo real que cualquier otro clasificador. También requiere una pequeña cantidad de datos de entrenamiento. El clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior. También es útil para datos dimensionales altos ya que la probabilidad de cada atributo se estima independientemente

2.1.3. Redes Neuronales

"Las redes neuronales son modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas." El modelo de redes neuronales. (s. f.). IBM. Recuperado 14 de abril de 2021, de https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model

Se han convertido en método de machine learning más conocidos, dentro de los usos que se le pueden dar a estos modelos son: reconocimiento de imágenes, traducción, análisis genético, reconocimiento de voz, predicción bursátil, generación de texto, entre otros usos.

La complejidad de estas redes se basa en la lógica que todo sistema grande funciona a través de la interacción de varios sistemas más pequeños y cuya parte más pequeña según su nombre lo indica es la neurona es una unidad básica de funcionamiento y el conjunto de neuronas es el que permite realizar las grandes acciones.

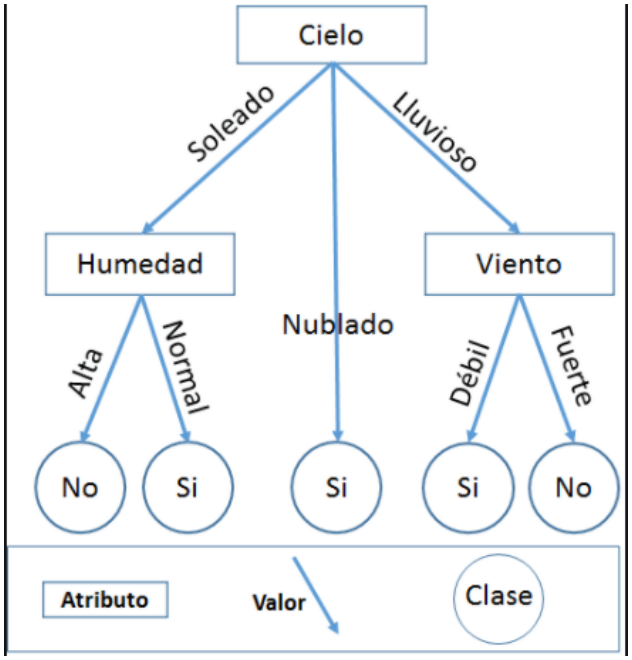


2.1.4. Arboles de Decisión

Este modelo de Machine Learning es un algoritmo de aprendizaje supervisado, es un modelo clásico de decisiones, como lo indica su nombre este modelo comienza desde un único nodo y se ramifica variables o alternativas. Cada hoja de la ramificación representa o constituye una clasificación o una decisión con el fin de encontrar la mejor opción.

Se puede interpretar también como un mapa de probables resultados de una cadena de pasos relacionadas entre sí, “Hay tres tipos diferentes de nodos: nodos de probabilidad, nodos de decisión y nodos terminales. Un nodo de probabilidad, representado con un círculo, muestra las probabilidades de ciertos resultados. Un nodo de decisión, representado con un cuadrado, muestra una decisión que se tomará, y un nodo terminal muestra el resultado definitivo de una ruta de decisión.” *Qué es un diagrama de árbol de*

decisión. (s. f.). Lucidchart. Recuperado 1 de abril de 2021, de https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision.



3. Resultados

3.1. Realizamos es la exportación de pandas el cual nos permite realizar manipulación y análisis de datos y LabelEncoder esta nos da la posibilidad de convertir datos strings en escalares.

```
# libreria pandas
import pandas as pd
from sklearn.preprocessing import LabelEncoder
```

3.2. Realizamos carga del archivo csv donde tenemos la data del proyecto, adicional se le configuran unas cosas a la lectura del csv por ejemplo le decimos que las cabeceras están desde el 0 y realizamos un encoding al archivo para que identifique el utf-8.

```
5   # Carga del archivo con configuraciones para que funcionen con utf-8
6   data = pd.read_csv('data.csv', header=0, encoding='unicode_escape')
```

3.3. Realizamos una limpieza de datos eliminando los datos nulos con dropna() un comando de pandas.

```
# Si tuvieran valores nulos
data = data.dropna()
```

3.4. Para un futuro proceso necesitamos almacenar la informacion de las columnas 5,6 y 7 por cual asignamso una variable con estos datos

```
# Asignar a variables valores de columnas especificas del archivo
toc = data.iloc[:, 4].values
animo = data.iloc[:, 5].values
obediencia = data.iloc[:, 6].values
```

3.5. Para poder utilizar el LabelEncoder necesitamos instanciarlo para poder utilizar mas propiedades de esta librería.

```
# Instancia de LabelEncoder
LabelEncoder_data = LabelEncoder()
```

3.6. Eliminamos unas columnas que nos crearían conflicto si las mantenemos por nuestro proceso de la data, el comando drop nos permite borrar varias columnas al tiempo solo tuvimos que indicar el nombre de la columna en la docuemntacion observamos que tambien se pueden borrar filas por lo que es importante especificarle cual queremos borrar.

```
# Borramos columnas que vamos a reemplazar
data.drop(['toc', 'animo', 'obediencia'],
          axis='columns', inplace=True)
```

3.7. Teníamos datos que no eran numéricos y que eran categorizados por lo tanto utilizamos la función de LaberEncoder fit_transform para que en cada columna consiguiera un patrón y volviera los datos numéricos y escalables.

```
# Escalar volver datos string en numeros
data['toc'] = LabelEncoder_data.fit_transform(toc)
data['animo'] = LabelEncoder_data.fit_transform(animo)
data['obediencia'] = LabelEncoder_data.fit_transform(obediencia)
```

3.8. Adicional encontramos que tomaba algunas columnas como tipo de dato flotante lo cual no resulta util en la lectura de nuestros datos por lo tanto casteamos algunas columnas con el type correspondient

```
# castear algunas columnas con el tipo de dato correspondiente
data = data.astype({
    "id_niño": int,
    "actividad": int,
    "categoria_actividad": int,
    "edad": int,
    "tiempo_rsp_seg": int,
    "fecha": 'datetime64'})
```

3.9. Decidimos que era prudente filtrar por fechas para obtener los datos que necesitamos en caso de que el archivo tenga fechas superiores a las que nos interesan.

```
# filtrar para que las fechas sean seleccionables
data = data[(data['fecha'] > '2020-01-01') & (data['fecha'] <= '2021-04-09')]
```

3.10. Nuestro proyecto necesita el filtro de rango de edades porque aplicamos el filtro a este campo.

```
# filtrar por edades
data = data[(data['edad'] > 4) & (data['edad'] <= 10)]
```

3.11. Por último, exportamos el archivo con la data limpia y con todos los filtros que consideramos necesarios a un csv nuevo.

```
# Exporta data a un archivo con pre-procesamiento de la data
data.to_csv('data_clean.csv', index=True, sep='|')
```

4. Bibliografía

- *Data and Statistics on Children's Mental Health | CDC. (2020, 15 junio). Centers for Disease Control and Prevention.* https://www.cdc.gov/childrensmentalhealth/data.html
- *Children's Mental Health Disorders - A Journey for Parents and Children*. (2014, 5 mayo). [Vídeo]. Youtube. https://youtube.com/watch?v=ewbD2Dw0NLo
- Bellido, R. M. (2019, 28 junio). *Psicología del preescolar: desarrollo de la conducta humana normal de los 3 a los 7 años de edad*. Repositorio Institucional Universidad Nacional. https://repositorio.unal.edu.co/handle/unal/38944
- Subdirección de Enfermedades No Trasmisibles Grupo Gestión Integrada para la Salud Mental. (2017, diciembre). *Boletín de salud mental Salud mental en niños, niñas y adolescentes* (N.º 4). https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/boletin-4-salud-mental-nna-2017.pdf

- Navarro-Pardo, E. (2012). *Desarrollo infantil y adolescente: trastornos mentales más frecuentes en función de la edad y el género*. Redalyc.org. https://www.redalyc.org/articulo.oa?id=72723439006
- https://aprendeia.com/algoritmos-de-clustering-agrupamiento-aprendizaje-no-supervisado/
- *https://www.clubdetecnologia.net/blog/2018/los-10-moadelos-mas-populares-de-inteligencia-artificial/*
- *http://posgrado.lapaz.tecnm.mx/uploads/archivos/TesisHdzCedano.pdf*
- *https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/mineria-de-datos-y-modelos-predictivos-descubriendo-patrones*
- *https://campusvirtual.ucc.edu.co/content/enforced/361842-01BOG_FINGE_PREG_01ISC_706164_2110_7864/7.%20Redes%20Neuronales%20SOM.pdf*
- *https://agenciab12.com/noticia/que-es-modelo-predictivo-como-aplica-negocio*
- *https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision.*
- *APUNTES DE CLASE DE PYTHON. (s. f.). APUNTES DE CLASE DE PYTHON. Recuperado 11 de abril de 2021, de https://campusvirtual.ucc.edu.co/content/enforced/361842-01BOG_FINGE_PREG_01ISC_706164_2110_7864/7.%20Redes%20Neuronales%20SOM.pdf*
- Mosquera, R. (s. f.). Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos. Scielo. https://scielo.conicyt.cl/scielo.php?pid=S071807642018000600153&script=sci_arttext