

PROYECTO FINAL MÁSTER
CLASIFICADOR DOCUMENTOS MÉDICOS HOPE
2020 - 2021

Ruben Vasallo Gonzalez

21 de julio de 2020

Índice general

1. Principal	2
1.1. Resumen	2
1.2. Abstract	2
1.3. Keywords	2
2. Introducción	3
2.1. Definición del proyecto	3
2.2. Estado del arte	3
3. Objetivos del Máster	4
3.1. Objetivo principal	4
3.2. Objetivos secundarios	4
4. Metodología	5
4.1. Extracción de los datos	5
4.1.1. Lectura de los datos	8
4.1.2. Conversión a formato columnar	9
4.2. Procesado de los datos	11
4.2.1. Análisis de los datos	11
4.2.2. Análisis de componentes principales	11
4.3. Enriquecimiento de los datos. Aproximación por Vecinos más próximos (K-NN)	12
4.4. Modelos Predictivos	13
4.4.1. Regresión logística ' <i>Logistic regression</i> '	13
4.4.2. Bosques Aleatorios ' <i>Random Forest</i> '	13
4.4.3. Maquinas de Vector Soporte ' <i>Support Vector Machines</i> '	13
5. Conclusiones	14
5.1. Resultados obtenidos	14
6. Bibliográfica	15
Índice de figuras	16
7. Anexos	17

Capítulo 1

Principal

1.1. Resumen

El proyecto nace de la necesidad de poder disponer de una manera sencilla e inmediata, artículos médicos catalogados según los síntomas de pacientes, pudiendo hacer un *ranking* de más o menos interés en función del *feedback* aportado por los profesiones médicos de artículos relacionados anteriormente.

1.2. Astract

TODO

1.3. Keywords

clasificador articulos medicos, PCA, KNN, Regresion Logistica

Capítulo 2

Introducción

2.1. Definición del proyecto

El proyecto de fin de máster que aquí se presenta nace de la necesidad por parte del *proyecto HOPE* de clasificar y recomendar resultados sobre estudios clínicos de confianza y que estén actualizados. En Internet existe muchísima información sobre medicina y salud y no siempre toda es de fiar.

El proyecto HOPE (que significa *Health Operations for Personalized Evidence* en inglés) nace de la necesidad de ayudar a los científicos médicos a encontrar la información que necesitan de la manera más rápida y fácil posible. Existe infinidad de información médica en Internet de miles de proyectos de investigación médica y esto hace que, muchas veces sea complicado encontrar la información sobre ensayos médicos para tratar información. En el ámbito de la medicina el tiempo perdido puede costar vidas y es un precio demasiado elevado a pagar, tanto a nivel económico como emocional.

Actualmente existen bases de datos de confianza en donde los científicos y el público en general puede buscar informes y ensayos sobre estudios clínicos desarrollados anteriormente, pero no siempre es fácil o rápido encontrar estos resultados.

El proyecto HOPE es un sistema basado en inteligencia artificial para identificar los datos claves de casos clínicos registrados en la Historia Clínica Electrónica, en base a los cuales realiza una búsqueda única por paciente para proporcionar al profesional de la salud recomendaciones de tratamientos, estudios de investigación, información para el paciente, todo en base a registros de fuentes científicas de información. En este proyecto, médicos de todo el mundo puede consultar en una base de datos informes médicos relacionados con los síntomas que puedan tener sus pacientes y ver que otros tratamientos han dado resultado. Todo y con eso, el sistema no siempre devuelve los artículos más relevantes o actualizados por lo que, no siempre la información consultada es útil.

En este ámbito, los profesionales médicos pueden valorar si la información recibida ha sido útil o no respecto a la búsqueda que han realizado, por lo que con ese *feedback*, se pretende mejorar el sistema actual complementándolo con un modelo clasificador capaz de ayudar al actual a entregar realmente los artículos útiles basándose en el *feedback* que los médicos dan al sistema.

2.2. Estado del arte

Recomendaciones que existen en el mundo actual

Capítulo 3

Objetivos del Máster

3.1. Objetivo principal

OP - Poder mostrar al usuario cuales son los artículos más útiles que pueden ayudar en el tratamiento del paciente, en base a los síntomas que este tiene, pudiendo realizar un *ranking* de mas interés a menos.

3.2. Objetivos secundarios

Para poder cumplir con el objetivo principal [OP1](#), desglosaremos los siguientes objetivos secundarios:

OS1 - Extraer la información de la base de datos y tratarla para quedarnos solo con la que consideramos valida.

OS2 - Hacer un análisis de componentes principales (estudio de que atributos son relevantes para alcanzar el objetivo).

OS3 - Enriquecer de los datos (*data augmentation*) prediciendo los resultados que no están indicados si son relevantes o no. Aproximación por Vecinos más próximos (*K-Nearest-Neighbor*).

OS4 - Predecir los resultados usando el algoritmo de aprendizaje supervisado para clasificación llamado Regresión logística "*Logistic regression*".

OS5 - Predecir los resultados usando el algoritmo de aprendizaje supervisado para clasificación llamado Bosques Aleatorios "*Random Forests*".

OS6 - Predecir los resultados usando el algoritmo de aprendizaje supervisado para clasificación llamado Máquinas de vector soporte "*Support Vector Machines*".

Capítulo 4

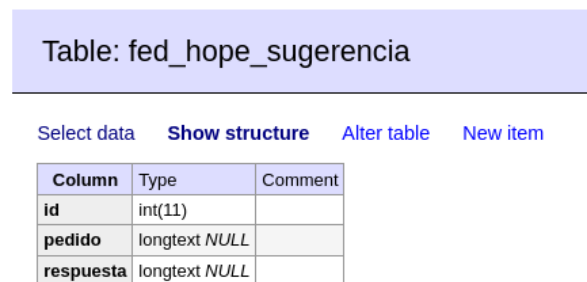
Metodología

Para cumplir con el [OP1](#) mostramos los pasos que hemos seguido para extraer y procesar los datos:

4.1. Extracción de los datos

El Origen de los datos se encuentra en una Base de datos SQL distribuida en dos tablas, que pasamos a detallar a continuación:

- Tabla *fed_hope_sugerencia*



The image shows a database management interface. At the top, a purple header bar contains the text 'Table: fed_hope_sugerencia'. Below this, there are four blue links: 'Select data', 'Show structure', 'Alter table', and 'New item'. The 'Show structure' link is highlighted. Below the links is a table with three columns: 'Column', 'Type', and 'Comment'. The table contains three rows of data:

Column	Type	Comment
id	int(11)	
pedido	longtext NULL	
respuesta	longtext NULL	

Figura 4.1: Visualización de los atributos de la tabla *fed_hope_sugerencia*

En esta tabla encontraremos la sugerencia que dio el programa HOPE en base a los parámetros que introdujo el usuario, almacenado en el atributo pedido y la respuesta que dio el programa, almacenado en el atributo respuesta. Todos los datos son almacenados en formato documento json.

Atributo pedido: Si analizamos el formato de datos que tenemos del atributo pedido observamos los siguientes atributos:

```
{
  "data": {
    "type": "emr--emr",
    "attributes": {
      "name": null,
      "affected_organ": "",
      "age": "75",
      "diagnostic_main": "FISTULA PERITONEAL",
      "gender": "male",
      "medical_history": "Paciente de 75 años con antecedentes de gastrectomía total por adenocarcinoma gástrico que intercurrió con eventración y posterior formación de fístula entero-atmosférica. Actualmente cursa postoperatorio de resección intestinal mu00e1s eventroplástica con colocación de malla. Al examen impresiona en regular estado general, lúcido, hemodinámicamente estable, sin signos de falla de bomba. Regular entrada de aires, rales crepitantes bilaterales. Abdomen blando, depresible, levemente doloroso a la palpación profunda. Herida cubierta por apósitos estériles. Catarsis positiva. Diuresis positiva.\n\nPROBLEMAS ACTIVOS:\n- POP resección intestinal mu00e1s eventroplástica estable, hemodinámicamente compensada. Refiere buena tolerancia al dolor. Afebril hace 72 hs. Cumple 4to día de tratamiento con piperacilina tazobactam por neumonía broncoaspiratoria con aislamiento de E.coli BLEE. Hemocultivos negativos. TAC de tórax y abdomen informa: Consolidación bibasal bilateral mu00e1s derrame pleural bilateral. Colección mu00e1s laminar posterior a ambos mu00e1s rectos de 10x2x0.4 cm, y otra colección mu00e1s en herida quirúrgica de pared de 10x2.8x1.2 cm. Se da aviso a cirujano tratante. Persiste con estado nauseoso, por lo que continúa con antieméticos reglados. \n\nEn aislamiento de contacto por germen multirresistente. \nSe da informe. Control evolutivo."
    }
  }
}
```

Podemos ver varios atributos haciendo referencia a los síntomas que consulta el profesional médico.

Atributo respuesta: En el atributo respuesta obtenemos entre otros datos, el listado de artículos científicos sugeridos relacionados con los síntomas descritos por el profesional médico. Esta respuesta es muy amplia pero entre todos los atributos, podemos observar un listado de identificadores de artículos, con sus fechas de revisión de estos, y unas palabras claves descriptivas para esos artículos.

Nota: Mostramos una pequeña parte del contenido de una observación del atributo respuesta.

```
{
  "data": {
    "type": "emr--emr",
    "id": "ef6d63fb-afe8-4650-8ab8-d4d75edd4fe5",
    "attributes": {
      "id": "376",
      "uid": "ef6d63fb-afe8-4650-8ab8-d4d75edd4fe5",
      "langcode": "es",
      "name": null,
      "status": true,
      "created": "1559052244",
      "changed": "1559052244",
      "affected_organ": null,
      "age": "75",
      "clinicaltrials": {
        "query": [],
        "trials": [],
        "diagnostic_main": "FISTULA PERITONEAL",
        "diagnostic_main_mesh_terms": null,
        "gender": "male",
        "history_date": null,
        "medical_history": "Paciente de 75 años con antecedentes de gastrectomía total por adenocarcinoma gástrico que intercurrió con eventración y posterior formación de fístula entero-atmosférica. Actualmente cursa postoperatorio de resección intestinal mu00e1s eventroplástica con colocación de malla. Al examen impresiona en regular estado general, lúcido, hemodinámicamente estable, sin signos de falla de bomba. Regular entrada de aires, rales crepitantes bilaterales. Abdomen blando, depresible, levemente doloroso a la palpación profunda. Herida cubierta por apósitos estériles. Catarsis positiva. Diuresis positiva."
      },
      "PROBLEMAS ACTIVOS": "POP resección intestinal mu00e1s eventroplástica estable, hemodinámicamente compensada. Refiere buena tolerancia al dolor. Afebril hace 72 hs. Cumple 4to día de tratamiento con piperacilina tazobactam por neumonía broncoaspiratoria con aislamiento de E.coli BLEE. Hemocultivos negativos. TAC de tórax y abdomen informa: Consolidación bibasal bilateral mu00e1s derrame pleural bilateral. Colección mu00e1s laminar posterior a ambos mu00e1s rectos de 10x2x0.4 cm, y otra colección mu00e1s en herida quirúrgica de pared de 10x2.8x1.2 cm. Se da aviso a cirujano tratante. Persiste con estado nauseoso, por lo que continúa con antieméticos reglados. \n\nEn aislamiento de contacto por germen multirresistente. \nSe da informe. Control evolutivo."
    },
    "medical_history_mesh_terms": [
      "Intestines",
      "Therapeutics",
      "Catharsis",
      "Wounds and Injuries",
      "Abdomen",
      "Respiratory Sounds",
      "Palpation",
      "Antiemetics",
      "Nausea",
      "Tazobactam",
      "Adenocarcinoma",
      "Fistula",
      "Pain Threshold",
      "Gastrectomy",
      "Signs and Symptoms",
      "Pleural Effusion",
      "Thorax",
      "Incisional Hernia",
      "Piperacillin",
      "Surgical Wound",
      "Diuresis",
      "Quarantine",
      "Muscles",
      "Blood Culture",
      "Tomography",
      "X-Ray Computed",
      "Pneumonia",
      "Pain",
      "medlineplus": {
        "query": [],
        "topics": [],
        "mesh_terms": [
          "Intestines",
          "Therapeutics",
          "Catharsis",
          "Wounds and Injuries",
          "Abdomen",
          "Respiratory Sounds",
          "Palpation",
          "Antiemetics",
          "Nausea",
          "Tazobactam",
          "Adenocarcinoma",
          "Fistula",
          "Pain Threshold",
          "Gastrectomy",
          "Signs and Symptoms",
          "Pleural Effusion",
          "Thorax",
          "Incisional Hernia",
          "Piperacillin",
          "Surgical Wound",
          "Diuresis",
          "Quarantine",
          "Muscles",
          "Blood Culture",
          "Tomography",
          "X-Ray Computed",
          "Pneumonia",
          "Pain",
          "aged",
          "male",
          "metastasis": null,
          "pubmed": {
            "query": {
              "db": "pubmed",
              "term": "u0022agedu0022[mesh] AND u0022maleu0022[mesh] AND (u0022Intestinesu0022[mesh] OR u0022Therapeuticsu0022[mesh] OR u0022Catharsisu0022[mesh] OR u0022Wounds and Injuriesu0022[mesh] OR u0022Abdomenu0022[mesh] OR u0022Respiratory Soundsu0022[mesh] OR u0022Palpationu0022[mesh] OR u0022Antiemeticu0022[mesh] OR u0022Nauseau0022[mesh] OR u0022Tazobactamu0022[mesh] OR u0022Adenocarcinomau0022[mesh] OR u0022Fistulau0022[mesh] OR u0022Pain Thresholdu0022[mesh] OR u0022Gastrectomyu0022[mesh] OR u0022Signs and Symptomsu0022[mesh] OR u0022Pleural Effusionu0022[mesh] OR u0022Thoraxu0022[mesh] OR u0022Incisional Hernia u0022[mesh] OR u0022Piperacillinu0022[mesh] OR u0022Surgical Woundu0022[mesh] OR u0022Diuresisu0022[mesh] OR u0022Quarantineu0022[mesh] OR u0022Musclesu0022[mesh] OR u0022Blood Cultureu0022[mesh] OR u0022Tomography, X-Ray Computedu0022[mesh] OR u0022Pneumoniau0022[mesh] OR u0022Painu0022[mesh])",
              "datetype": "edit",
              "retmax": 10,
              "sort": "relevance",
            },
            "articles": [
              {
                "id": "27395425",
                "title": "Indications and Results of Reconstructive Techniques with Flaps Transposition in Patients Requiring Complex Thoracic Surgery: A 12-Year Experience.",
                "abstract": {
                  "attributes": {
                    "Label": "BACKGROUND",
                    "value": "Flap transposition is an infrequent but far from exceptional thoracic surgical procedure. The aim of this retrospective study was to report our experience in a referral unit of general thoracic surgery analyzing the early results after flap transposition."
                  },
                  "attributes": {
                    "Label": "METHODS",
                    "value": "We retrospectively analyzed the clinical records, surgical notes, and postoperative results of a cohort of patients who underwent flap transposition in our unit from November 2000 to February 2013."
                  },
                  "attributes": {
                    "Label": "RESULTS",
                    "value": "Overall, a surgical approach adopting flap reconstruction techniques was performed in 81 patients (54 males, 27 females) with a median age of 62u00a0u00a0years (range 20-87). Flap transposition was necessary to reconstruct chest wall after resection for malignancy (27 patients), to repair intrathoracic viscera perforation (15 patients), and to fill residual cavities secondary to pulmonary resection (19 patients). A single muscle flap was transposed in most of cases (64 out of 70u00a0u00a0cases), while in the remaining 17 cases"
                  }
                }
              ]
            }
          }
        }
      }
    }
  }
}
```

Figura 4.2: Ejemplo de contenido del atributo respuesta de una observación

- Tabla *fed_hope_sugerencia_feedback*

Table: fed_hope_sugerencia_feedback

Select data Show structure Alter table New item

Column	Type	Comment
id	int(11)	
articulo	varchar(255) NULL	
utilidad	int(11) NULL	
comentario	varchar(255) NULL	
fed_hope_sugerencia_id	int(11) NULL	

Figura 4.3: Visualización de los atributos de la tabla *fed_hope_sugerencia_feedback*

En esta tabla encontraremos la opinión *feedback* (que utilidad ha tenido la información por parte del usuario) de la información recibida dado un artículo en concreto en una búsqueda en concreto. Esta información se relaciona con la tabla *fed_hope_sugerencia* a través del atributo *fed_hope_sugerencia_id*.

En esta tabla esta representada la opinión *feedback* del usuario en el atributo *utilidad*, que denota un valor 0 para los artículos que han sido poco útiles respecto a la búsqueda realizada y 1 para los artículos que si han sido útiles.

4.1.1. Lectura de los datos

El primer paso sera extraer de la base de datos *MySQL* el dataset de datos. Para realizar esta acción nos ayudaremos de las librerías *sqlalchemy* y *pymysql* programadas en lenguaje *python* que nos permitirá acceder a la información almacenada en una base de datos *MySQL* y devolvérsola en formato *dataframe*, un formato que nos permite entre otras cosas, realizar transformaciones de los datos para conseguir nuestro objetivo final.

Este formato es interpretable por la librería *pandas* y *numpy*, dos librerías programadas en lenguaje *python*, muy comunes en el ámbito de la ciencia del dato, que nos facilitara entre otras cosas, poder hacer operaciones matemáticas con los datos de manera eficiente.

Import data from DB.

```
In [1]: # pip install pymysql
from sqlalchemy import create_engine
import pymysql
import pandas as pd
import numpy as np

In [2]: dbConnectionURL = 'mysql+pymysql://root:hope@mysql-master/hope'
dbConnection = create_engine(dbConnectionURL)

df = pd.read_sql('SELECT id, pedido, respuesta FROM fed_hope_sugerencia', con=dbConnection)

In [3]: df.head(10)
```

Out[3]:

	id	pedido	respuesta
0	29	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'ef6d63fb-afe8...
1	30	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'0b8a1cc8-ce17...
2	31	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'25733e18-3245...
3	32	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'40320232-7510...
4	33	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'f686d89e-fc8e...
5	34	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'d94d7c78-9941...
6	35	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'0a14cc6b-af7b...
7	36	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'245bb87d-b52c...
8	37	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'fad05206-04f6...
9	38	{'data':{'type':'emr--emr','attributes':{'name'...	{'data':{'type':'emr--emr','id':'a0f8dabe-a795...

Figura 4.4: Lectura de los datos de Mysql

4.1.2. Conversión a formato columnar

Para poder trabajar con los datos, necesitaremos que estos estén en formato columnar (tabla relacional) por lo que necesitaremos convertir los datos de estos *json* en tablas relacionales (A esta acción se le conoce como *flattening* o aplanar).

Este paso consiste en coger cada uno de los atributos que tiene el *json* y convertirlos en columnas de una tabla, añadiendo los valores. Si el *json* tiene varios niveles, este proceso añadirá tantas columnas como niveles tenga el *json*, siempre que todas las observaciones del *json* tengan el mismo formato. Este caso se nos cumple para las observaciones del atributo Pedido. No es así para las observaciones del atributo respuesta en el que tendremos que hacer un tratamiento especial que detallaremos posteriormente.

Para poder realizar esta acción, primero tendremos que solucionar un problema de tratamiento de texto. Los datos que se nos proporcionan, contienen caracteres que informan de un salto de línea o tabulador. Estos caracteres pueden ser mal interpretados a la hora de leer los datos de los documentos en formato *json* por lo que los eliminaremos de los datos (substituyendo-los por espacios vacíos).

Replace illegal chars in json

```
In [4]: mapping = {'\r\n': ' ', '\n': ' ', '\t': ' ', r'\\.': ' ',}
df = df.replace({'pedido': mapping, 'respuesta': mapping}, regex=True)
df.head(10)
```

```
Out[4]:
```

	id	pedido	respuesta
0	29	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"ef6d63fb-afe8..."
1	30	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"0b8a1cc8-ce17..."
2	31	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"25733e18-3245..."
3	32	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"40320232-7510..."
4	33	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"f686d89e-fc8e..."
5	34	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"d94d7c78-9941..."
6	35	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"0a14cc6b-af7b..."
7	36	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"245bb87d-b52c..."
8	37	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"fad05206-04f6..."
9	38	{"data":{"type":"emr-emr","attributes":{"name..."	{"data":{"type":"emr-emr","id":"a0f8dabe-a795..."

Figura 4.5: Eliminación de caracteres erróneos.

• *Flattening* del atributo pedido

Para realizar la acción de *flattening* en el atributo pedido, nos ayudaremos de la funcionalidad *json_normalize* del paquete pandas que realiza esta acción.

Flattening JSON

```
In [5]: import ast
import json
from pandas import read_json, json_normalize #package for flattening json in pandas df
#https://stackoverflow.com/questions/39899005/how-to-flatten-a-pandas-dataframe-with-some-columns-as-json

pd.options.display.max_columns = None
#pd.options.display.max_rows = None

In [6]: # Flatterin column "Pedido"
pedidosData = json_normalize(df['pedido'].apply(json.loads).tolist()).add_prefix('pedido.')
pedidosData

Out[6]:
```

	pedido.data.type	pedido.data.attributes.name	pedido.data.attributes.affected_organ	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	pedido
0	emr--em	None		75	FISTULA PERITONEAL	
1	emr--em	None		31	REHABILITACION NEUROLOGICA	
2	emr--em	None		76	INSUFICIENCIA CARDIACA	
3	emr--em	None		75	FISTULA PERITONEAL	
4	emr--em	None		31	REHABILITACION NEUROLOGICA	
...
119	emr--em	None		74	DIFICULTAD RESPIRATORIA	
120	emr--em	None		48	REHABILITACION NEUROLOGICA	
121	emr--em	None		40	REHABILITACION NEUROLOGICA	
122	emr--em	None		43	TEP	
123	emr--em	None		37	DOLOR ABDOMINAL	

124 rows x 7 columns

Figura 4.6: *Flattening* del atributo pedido.

• *Flattening* del atributo respuesta

Debido a que la información almacenada en el documento *json*, en el atributo respuesta es muy compleja (debido a que esta contiene diferentes documentos con diferentes niveles de información) como se puede apreciar en el apartado X, no podemos aplanar la información directamente como hemos hecho con el atributo pedido. Por lo que tenemos que analizar que información nos interesa recoger para enriquecer el dataset de datos.

Después de analizar el documento, y ayudarnos del conocimiento del cliente, vemos que los atributos más interesantes son los que hacen referencia al identificador del artículo, las palabras claves asociadas al artículo por parte de la api pubmed y el mes y año de la revisión del artículo.

```
In [7]: # Flattenin column "respuesta"

def get_articles_from_respuesta(ld):
    jsonData = json.loads(ld)
    pubmedKeys = jsonData['data']['attributes']['pubmed_mt_opt']
    if pubmedKeys is None: pubmedKeys = []

    articles = list(jsonData['data']['attributes']['pubmed']['articles'])
    articlesIDs = []
    articlesRevisedYear = []
    articlesRevisedMonth = []
    for article in articles:
        articlesIDs.append(article['id'])
        articlesRevisedYear.append(article['revisedDate']['Year'])
        articlesRevisedMonth.append(article['revisedDate']['Month'])

    return dict({
        'articles': articlesIDs,
        'articlesRevisedYear': articlesRevisedYear,
        'articlesRevisedMonth': articlesRevisedMonth,
        'pubmed_keys': ','.join(pubmedKeys)
    })

respuestaData = json_normalize(df['respuesta'].apply(get_articles_from_respuesta).tolist()).add_prefix('respuesta.')
respuestaData
```

Out[7]:

	respuesta.articles	respuesta.articlesRevisedYear	respuesta.articlesRevisedMonth	respuesta.pubmed_keys
0	[27395425, 28560554, 28641726, 26245344, 28942...	[2018, 2018, 2017, 2016, 2018, 2014, 2018, 201...	[01, 04, 12, 12, 06, 06, 09, 04, 01, 04]	Intestines,Therapeutics,Catharsis,Wounds and I...
1	[30210096, 27617939, 27210858, 26412482, 25487...	[2019, 2017, 2017, 2017, 2016, 2016, 2019, 201...	[03, 04, 08, 06, 06, 09, 02, 03, 01, 11]	Back,Wounds and Injuries,Catheterization,Rest,...
2	[21067951, 27616270, 27532500, 28426556, 27495...	[2011, 2017, 2017, 2019, 2017, 2009, 2017, 201...	[03, 07, 05, 01, 07, 03, 06, 05, 03, 03]	Heart Murmurs,Intestines,Lactic Acid,Therapeut...
3	[30179656, 28641726, 28694230, 27796647, 28867...	[2019, 2017, 2018, 2017, 2017, 2017, 2015, 201...	[03, 12, 05, 02, 11, 05, 08, 06, 01, 08]	Intestines,Therapeutics,Catharsis,Lower Extrem...
4	[29787536, 24840763, 28273653, 26836795, 26409...	[2019, 2014, 2017, 2016, 2016, 2019, 2013, 201...	[05, 10, 11, 11, 05, 04, 03, 09, 02, 12]	Abdomen,Catheterization,Headache,Diuresis,Extr...
...
119	[28641726, 30179656, 28694230, 27796647, 28867...	[2017, 2019, 2018, 2017, 2017, 2017, 2017, 201...	[12, 03, 05, 02, 11, 05, 09, 09, 01, 03]	Extremities,Catharsis,Tazobactam,Abdomen,Oxyge...
120	[27128826, 30336861, 30226191, 29371130, 29587...	[2017, 2019, 2019, 2018, 2018, 2019, 2019, 201...	[04, 01, 09, 10, 08, 01, 06, 08, 11, 04]	Catharsis,Abdomen,Lung
121	[30595510, 21554494, 26465238, 26875969, 30056...	[2019, 2012, 2016, 2016, 2019, 2019, 2016, 201...	[03, 04, 09, 08, 07, 10, 10, 05, 10, 06]	Abdomen,Wounds and Injuries,Lung,Stroke,Aphasi...
122	[30081165, 30629460, 26220984, 25749653, 28545...	[2018, 2019, 2016, 2016, 2018, 2020, 2015, 201...	[12, 03, 06, 04, 12, 02, 02, 09, 05, 04]	Extremities,Catharsis,Thromboembolism,Foramen ...
123	[30662053, 29879068, 26849395, 31061178, 31223...	[2019, 2018, 2016, 2019, 2019, 2019, 2018, 201...	[02, 06, 12, 12, 07, 03, 12, 05, 04, 04]	Fever,Catharsis,Infections,Abdominal Pain,Abdo...

124 rows x 4 columns

Figura 4.7: *Flattening* del atributo respuesta.

4.2. Procesado de los datos

TODO

4.2.1. Análisis de los datos

TODO

4.2.2. Análisis de componentes principales

TODO

4.3. Enriquecimiento de los datos. Aproximación por Vecinos más próximos (K-NN)

TODO

4.4. Modelos Predictivos

4.4.1. Regresión logística '*Logistic regression*'

TODO

4.4.2. Bosques Aleatorios '*Random Forest*'

TODO

4.4.3. Maquinas de Vector Soporte '*Support Vector Machines*'

TODO

Capítulo 5

Conclusiones

5.1. Resultados obtenidos

TODO

Capítulo 6

Bibliográfica

Índice de figuras

4.1. Visualización de los atributos de la tabla <i>fed_hope_sugerencia</i>	5
4.2. Ejemplo de contenido del atributo respuesta de una observación	6
4.3. Visualización de los atributos de la tabla <i>fed_hope_sugerencia_feedback</i>	7
4.4. Lectura de los datos de Mysql	8
4.5. Eliminación de caracteres erróneos.	9
4.6. <i>Flattening</i> del atributo pedido.	10
4.7. <i>Flattening</i> del atributo respuesta.	11

Capítulo 7

Anexos