Clasificador Documentos Médicos HOPE

Trabajo Final De Máster, Área 2

Rubén Vasallo González

January 2, 2021

UOC

Índice

Introducción

Introducción

- Ayudar a los profesionales sanitarios a encontrar referencias bibliográficas adaptadas y personalizadas al paciente.
- Ayudar al proyecto HOPE a mejorar su algoritmo de Inteligencia Artificial (NLP).

Objetivos

- **(OP)** Recomendar al profesional sanitario las referencias bibliográficas actuales útiles y personalizadas pudiendo realizar una clasificación (ranking) de más interés a menos.

Objetivos

- **(OP)** Recomendar al profesional sanitario las referencias bibliográficas actuales útiles y personalizadas pudiendo realizar una clasificación (ranking) de más interés a menos.
- *(OS)* Analizamos el Conjunto de datos y componentes.
- *(OS)* Enriquecemos el conjunto de datos.
- *(OS)* Analizamos 3 posibles modelos predictivos.
- *(OS)* Realizamos la recomendación final

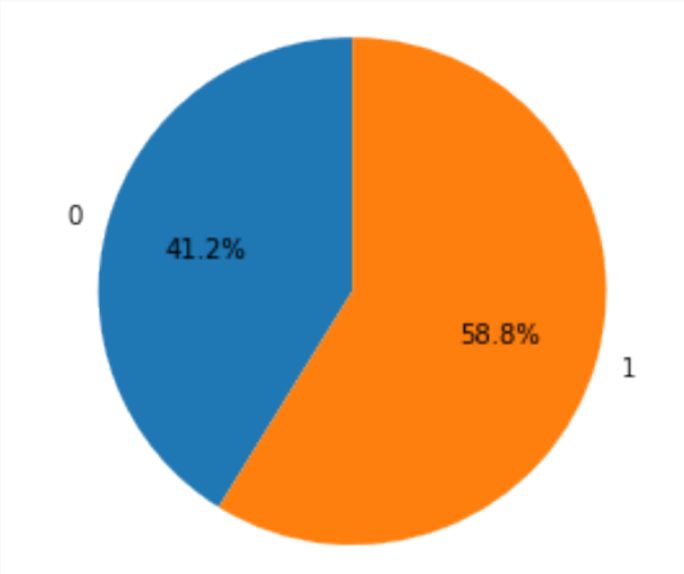
Metodología

Conjunto de Datos

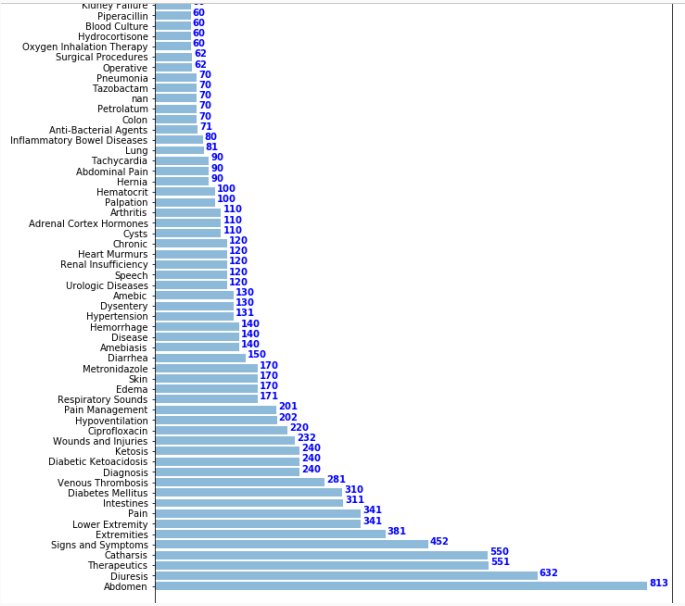
- **age**: 75,86,40,...
- **diagnostic_main**: Fistula Peritoneal, Insuficiencia Respiratoria,...
- **gender**: male
- **artículo**: 28694230,28805236,...
- **articlesRevisedYear**: 2018,2017,2016,...
- **articlesRevisedMonth**: 4,12,6,9,...
- **pubmed_keys**: (Abdomen, Adenocarcinoma, Antiemetics, Blood), (Abdomen, Analgesics, Bone, Catharsis), (Abdomen, Anti-Bacterial Agents, Diuresis),...
- **utilidad**: 0,1,NA

Limitaciones encontradas

- Poco volumen de información (51 obs. con atrb. *utilidad* informado).
- Atributo a predecir sin la información suficiente para poder hacer un *ranking*.
- Conjunto de datos Sesgado.



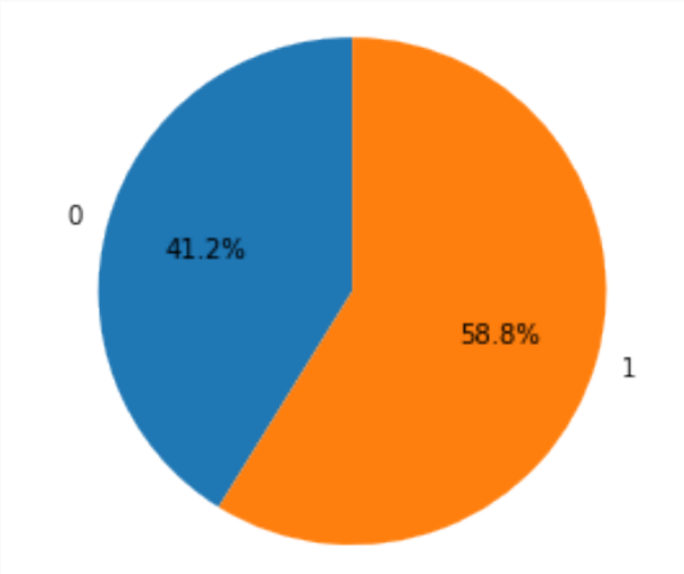
(a) Atributo *utilidad*



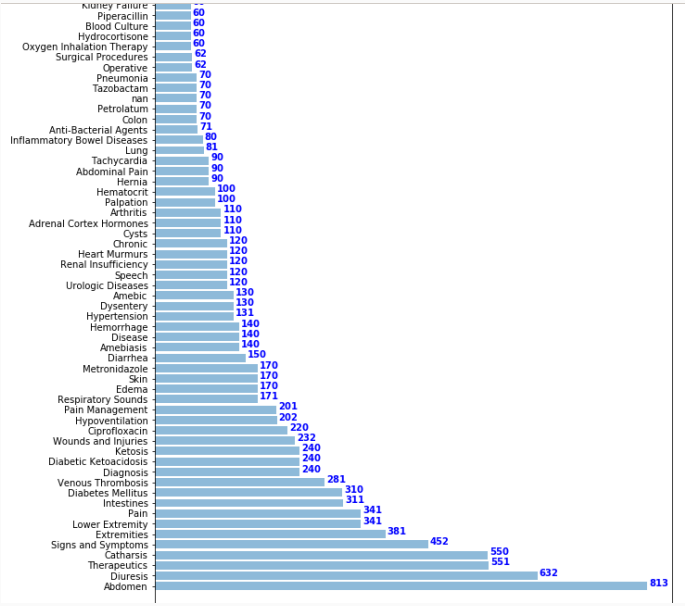
(b) Atributo *pubmed_keys*

Limitaciones encontradas

- Poco volumen de información (51 obs. con atrb. *utilidad* informado).
- Atributo a predecir sin la información suficiente para poder hacer un *ranking*.
- Conjunto de datos Sesgado.



(a) Atributo *utilidad*



(b) Atributo *pubmed_keys*

Análisis de componentes principales (PCA)

- **Conjunto 1**: Dataset Completo
- **Conjunto 2**: Dataset Completo pero añadiendo el mes y año del artículo y cogiendo solo las observaciones que se ha informado el atributo utilidad.
- **Conjunto 3**: Dataset conjunto 2 pero eliminando los atributos *gender* y artículo y se expande el atributo *pubmed_keys*

Análisis de componentes principales (PCA)

- Se Transformo todos los atributos Categóricos (texto) a Continuos (números continuos).
- Se estandarizó los valores a un rango de entre 1 y -1.
- Para el entrenamiento de los modelos predictivos, se dividió el conjunto en dos grupos. 1 Grupo de entrenamiento con el 75% de observaciones. 2 Grupo para la validación con el 25% de observaciones.

Resultados del análisis de componentes principales (PCA)

- Conjunto 1: Con solo **3 atributos**, el modelo es capaz de **explicar (predecir) el 95%** de las observaciones.

```
pd.DataFrame(pca.components_,columns=features,index = ['PC1','PC2','PC3'])
```

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	pedido.data.attributes.gender	respuesta.pubmed_keys	articulo
PC1	-0.050066	0.705007	-1.110223e-16	0.705462	0.052744
PC2	0.760475	0.071805	-1.318390e-16	0.030411	-0.644668
PC3	-0.630542	-0.130930	-1.110223e-16	0.142297	-0.751682

- Conjunto 2: Con **5 atributos**, el modelo es capaz de **explicar (predecir) el 97%** de las observaciones.

```
pd.DataFrame(pca.components_,columns=features,index = ['PC1','PC2','PC3','PC4','PC5'])
```

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	pedido.data.attributes.gender	respuesta.articlesRevisedYear	respuesta.articlesRevisedMont
PC1	-0.094054	0.603912	-1.110223e-16	-0.365651	0.33837
PC2	0.135318	0.327845	-5.273559e-16	0.362710	-0.55606
PC3	-0.891259	-0.174252	-5.828671e-16	-0.236812	-0.13360
PC4	0.382620	-0.115174	-1.929013e-15	-0.621075	0.16245
PC5	0.098479	0.051453	3.774758e-15	-0.540677	-0.72925

PC1 => diagnostic_main PC2 => pubmed_keys/Year PC3 => pubmed_keys PC4 => age PC5 => articulo

Resultados del análisis de componentes principales (PCA)

- Conjunto 3: Con solo **4 atributos**, el modelo es capaz de **explicar (predecir) el 90%** de las observaciones.

```
pd.DataFrame(pca.components_,columns=features,index = ['PC1','PC2','PC3','PC4'])
```

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	respuesta.articlesRevisedYear	respuesta.articlesRevisedMonth	respuesta.pubmed_keys
PC1	0.202186	0.688327	-0.236152	-0.080358	0.650462
PC2	-0.299028	-0.000864	-0.648047	0.698266	-0.055149
PC3	-0.908866	-0.006936	0.153411	-0.221720	0.318152
PC4	-0.015928	-0.145721	-0.702451	-0.673071	-0.179024

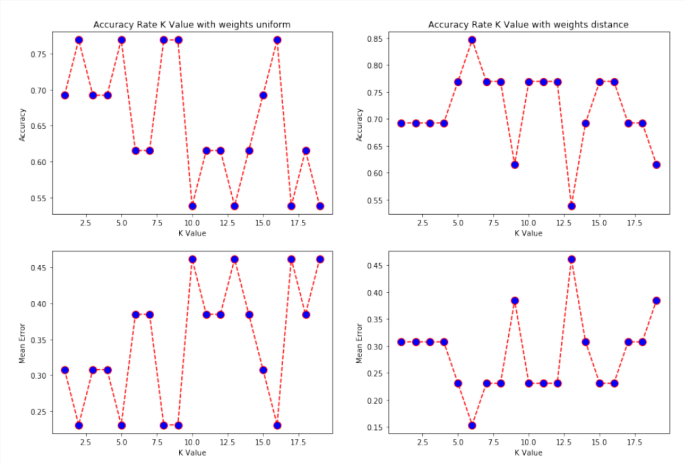
PC1 => diagnostic_main / pubmed_keys PC2 => articlesRevisedMonth PC3 => pubmed_keys PC4 => age

Intentando enriquecer los datos (K-Nearest-Neighbor)

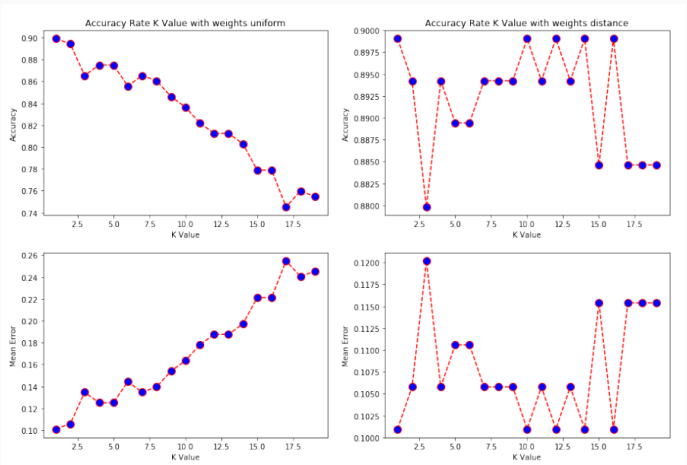
- **Conjunto 1**: Dataset Completo cogiendo solo las observaciones que se ha informado el atributo utilidad.
- **Conjunto 2**: Dataset conjunto 1 pero eliminando los atributos *gender* y artículo y se expande el atributo *pubmed_keys*

Resultados del enriquecimiento (K-Nearest-Neighbor)

- **Conjunto 1**: $\mathbf{K = 6}$ utilizando el calculo de la distancia '**distance**', con un porcentaje de acierto del **85%**.
- **Conjunto 2**: $\mathbf{K = 1}$ utilizando el calculo de la distancia '**uniform**', con un porcentaje de acierto del **90%**.



(a) Conjunto 1



(b) Conjunto 2

Análisis Modelos Predictivos

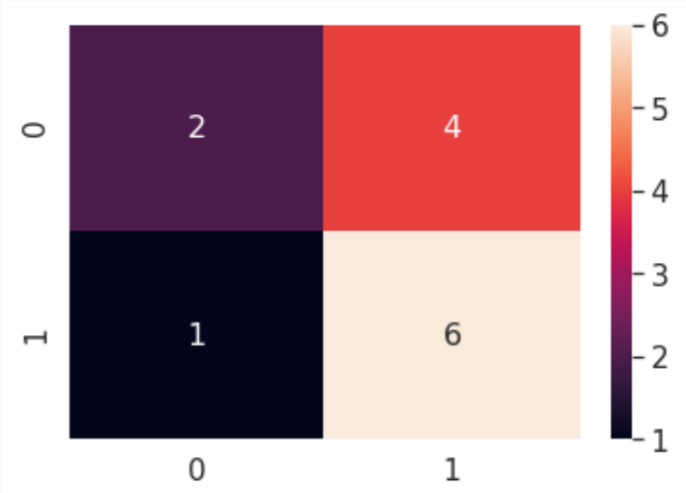
Conjuntos de datos para el estudio de los modelos

- **Conjunto 1**: Dataset Completo cogiendo solo las observaciones que se ha informado el atributo utilidad.
- **Conjunto 2**: Dataset conjunto 1 pero eliminando los atributos *gender* y artículo y se expande el atributo *pubmed_keys*

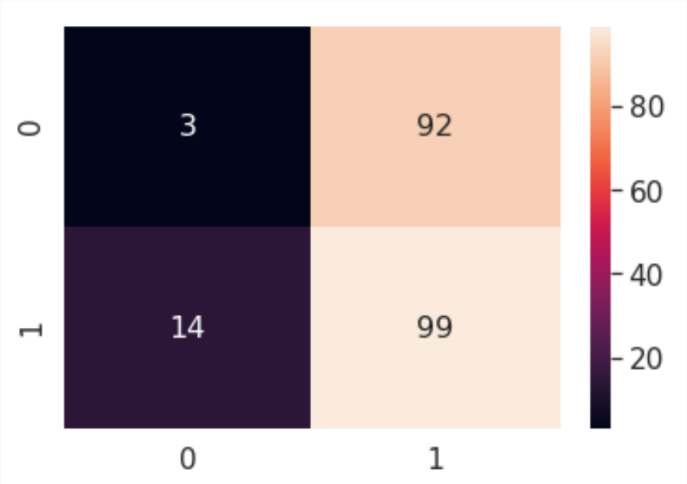
Nota: Se utilizaron los mismos conjuntos para los 3 modelos y se aplican las mismas transformaciones mencionadas anteriormente.

Resultados del Modelo 1 (Regresión logística)

- **Conjunto 1**: Precisión del **65%** sobre el conjunto de validación.
- **Conjunto 2**: Precisión del **49%** sobre el conjunto de validación.



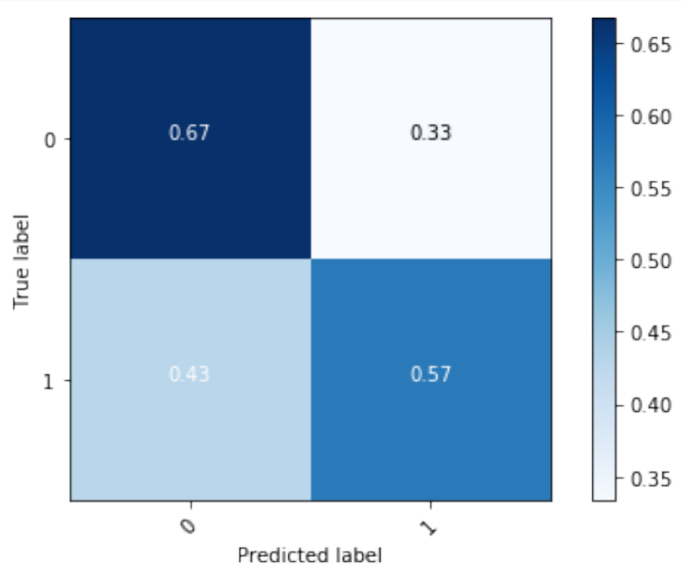
(a) Conjunto 1



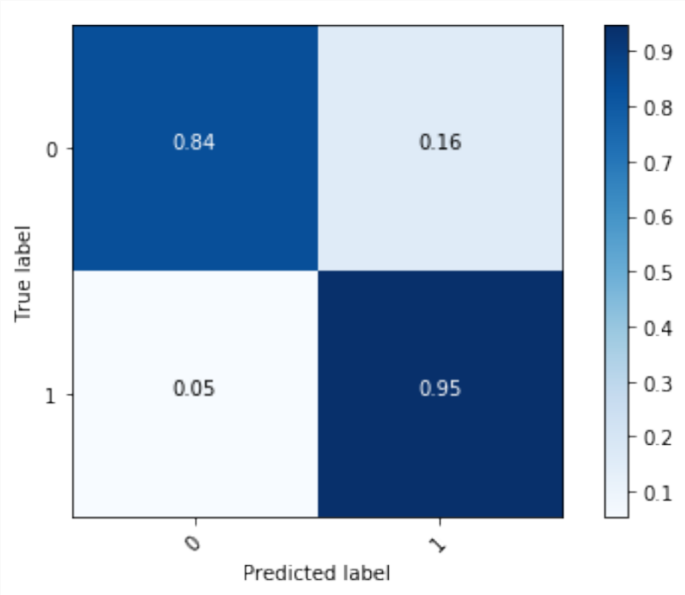
(b) Conjunto 2

Resultados del Modelo 2 (Random Forests)

- **Conjunto 1**: **N_Estimator: 40** con un porcentaje de acierto del **61%**.
- **Conjunto 2**: **N_Estimator: 10** con un porcentaje de acierto del **89%**.



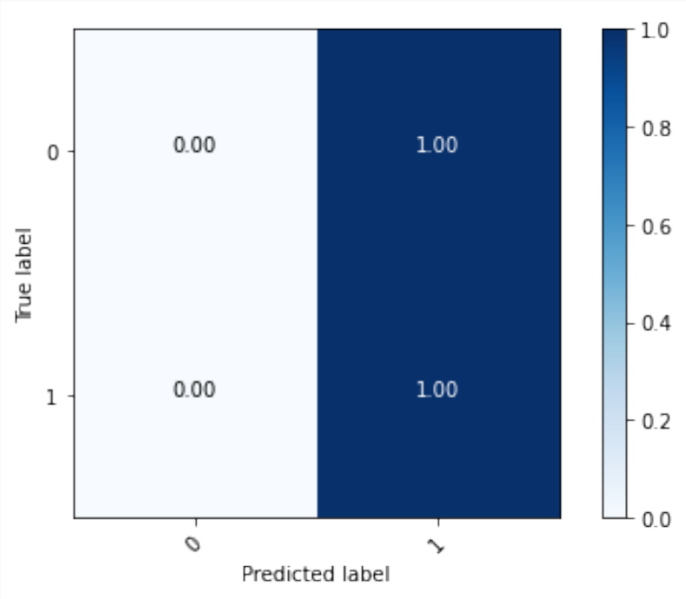
(a) Conjunto 1



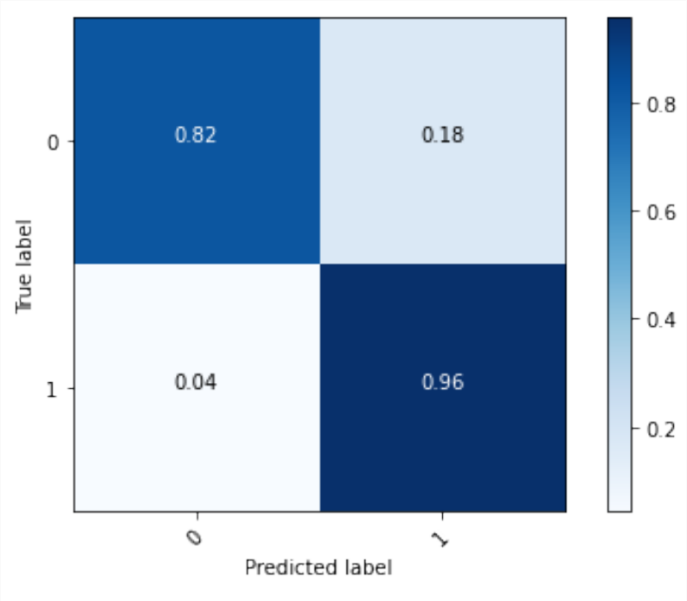
(b) Conjunto 2

Resultados del Modelo 3 (Support Vector Machines)

- **Conjunto 1**: **kernel: radial (rbf)** con un porcentaje de acierto del **54%**.
- **Conjunto 2**: **kernel: radial (rbf)** con un porcentaje de acierto del **89%**.



(a) Conjunto 1



(b) Conjunto 2

Conclusiones y Trabajos futuros

Comparando los Modelos

Modelos	Logit	**Bosques Aleatorios**	SVM
Conjunto 1	65.78%	61.53%	53.85%
Conjunto 2	49.03%	**89.9%**	89.42%

Trabajos futuro

- Mejorar el modelo actual.
  - Mejorar el actual modelo de Bosques aleatorios añadiendo más observaciones.
  - Valorar si otros modelos predictivos tienen mejor resultado.
  - Valorar si se altera la importancia de los atributos relevantes.
- Entrenar un modelo capaz de realizar un ranking de utilidad.

Gracias por esta oportunidad

¿Preguntas, dudas?

rvasallo@uoc.edu