# ANÁLISIS DESCRIPTIVO, INFERENCIA Y APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA BASES DE DATOS SOBRE SALUD Y SEGURIDAD EN EL TRABAJO PARA COLOMBIA

Hernández. A. Jairo, Presentado a: Ing. Julio Ochoa

**Resumen – Se realiza una búsqueda preliminar en bases de datos abiertas del gobierno, del Departamento Nacional de Estadística y demás organismos de carácter público y privado que puedan suministrar información sobre registros de accidentabilidad y enfermedad laboral, así, en primera instancia se realiza un análisis preliminar sobre consolidados entre los años 2017-2018 , para luego abordar un dataset más robusto abierto al público por Positiva seguros.**

## I. INTRODUCCION

La carencia de información representativa sobre parámetros que permitan establecer relaciones causales entre variables, son ecosistemas que no han permitido la evolución progresiva de técnicas cuantitativas con base estadística, más aún en países de Latinoamérica como Colombia, donde las fuentes de información secundarías y/o terciarías flaquean al indagar por información fehaciente y suficiente de la población local; no obstante, algunas fuentes en línea han estado actualizando su repositorio, situación la cual llevo a una búsqueda empírica de información, que finalmente resultó reflejada en el uso de dos *datasets*, en primera instancia, se elabora un análisis descriptivos de las variables dispuestas en la tabla ligada al siguiente link https://www.datos.gov.co/Trabajo/Consolidado-estad-sticas-accidentes-y-enfermedades/8d43-28hk, la cual retorna un consolidado de accidentes y enfermedades laborales - Del Grupo de Promoción y prevención de la Dirección de Riesgos de Laborales del Ministerio del Trabajo para el año 2017-2018:

```
df.head(3)
```

| | Mes | Año | Empresas Afiliadas | Afiliados Dependientes | Afiliados Independientes | Total Afiliados | Presuntos Accidentes de Trabajo | Accidentes de Trabajo Calificados | Er |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ENERO | 2017 | 710749 | 9193033 | 424847 | 9617880 | 53515 | 49451 | |
| 1 | FEBRERO | 2017 | 721453 | 9465116 | 544022 | 10009138 | 56976 | 54574 | |
| 2 | MARZO | 2017 | 727823 | 9612486 | 605740 | 10218226 | 65718 | 58818 | |

Fig. 1. Vista preliminar del archivo.csv

Posteriormente se procede a la evaluación de la consistencia de los datos (duplicados, valores nulos, atípicos, invalidos…), una vez realizado este procedimiento, ya es posible realizar operaciones con la información de entrada;

| | count | mean | std | |
|---|---|---|---|---|
| Año | 24.0 | 2.017500e+03 | 0.510754 | |
| Empresas Afiliadas | 24.0 | 7.746118e+05 | 37652.018327 | |
| Afiliados Dependientes | 24.0 | 9.593749e+06 | 201284.755249 | 9 |
| Afiliados Independientes | 24.0 | 7.527135e+05 | 142328.398462 | |
| Total Afiliados | 24.0 | 1.034646e+07 | 305586.667273 | 9 |
| Presuntos Accidentes de Trabajo | 24.0 | 5.831162e+04 | 4423.635381 | |
| Accidentes de Trabajo Calificados | 24.0 | 5.420838e+04 | 3648.537477 | |
| Presuntas Enfermedades Laborales | 24.0 | 1.636917e+03 | 913.810699 | |
| Enfermedades Laborales Calificadas | 24.0 | 8.374167e+02 | 113.609138 | |
| Muertes Accidentes de Trabajo Reportadas | 24.0 | 7.070833e+01 | 10.071394 | |
| Muertes Accidentes de Trabajo Calificadas | 24.0 | 4.687500e+01 | 10.613005 | |
| Nueva Pensión Invalidez Pagada Accidentes de Trabajo | 24.0 | 3.691667e+01 | 9.178977 | |
| Nueva Pensión Invalidez Pagada Enfermedad Laboral | 24.0 | 6.666667e+00 | 2.869126 | |
| Incapacidad Permanente Parcial Pagada por Accidente de Trabajo | 24.0 | 8.194167e+02 | 85.399784 | |
| Incapacidad Permanente Parcial Pagada por Enfermedad Laboral | 24.0 | 4.322500e+02 | 58.796074 | |

Fig. 1. Vista preliminar de un análisis descriptivo básico para cada atributo, es notable la alta desviación en la categoría 'Total de afiliados'

Así, se procede a continuar a la visualización y agregación de los datos para describir relaciones entre estos

F. A. Arthur Keras application a networking, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Computer Sciencel Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

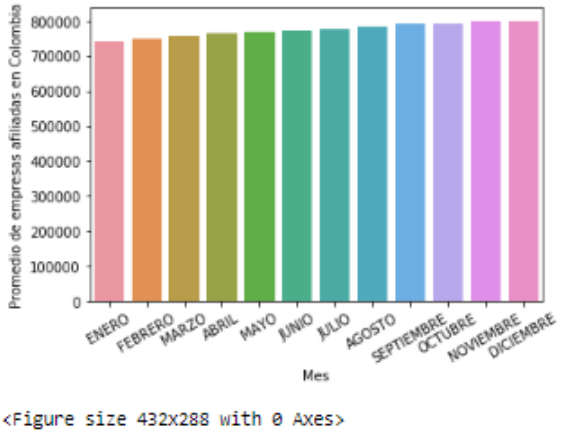


Fig. 3. Promedio de empresas afiliadas en Colombia, por mes, para año 2017-2018, agregado

Homólogamente se continua la usqueda de relaciones significativas entre los datos, que puedan ser esclarecidas a través de herramientas gráficas:

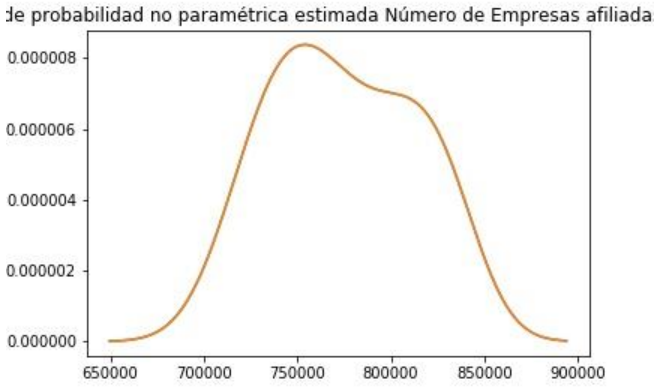


Fig. 3. Densidad de probabilidad no paramétricas determinada para el número de empresas afiliadas, es de apreciar el sesgo y apuntalamiento asimétrico de la distribución hacía valores por debajo de la media

Determinar si existe y decir así en cuanto fue el aumento porcentual entre años respecto al promedio de empresas afiliadas en Colombia:

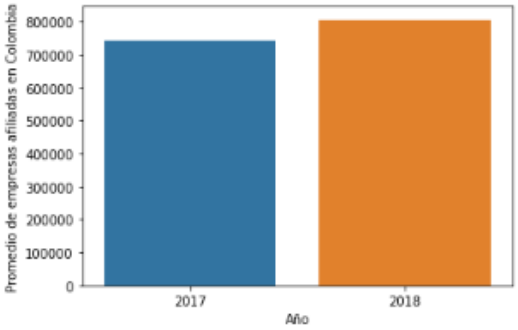


Fig. 4. Existe un incremento entre el promedio de afiliados en Colombia, cuantificado alrededor de 8,57 %

Así mismo se puedes establecer variaciones mensuales entre meses consecutivos

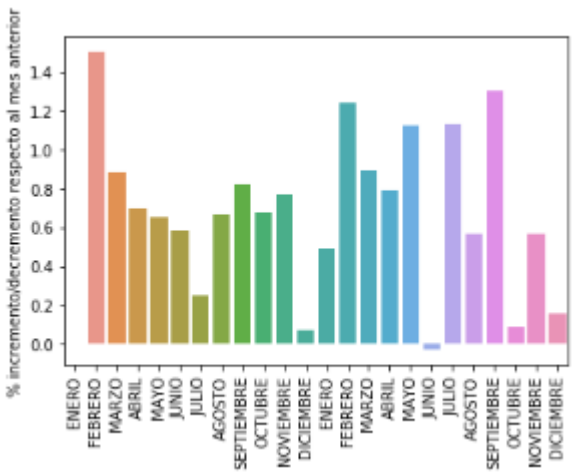


Fig. 5. La variación máxima significativa se da precisamente entre meses consecutivos iniciales Enero-Febrero 2017

Otro acercamiento práctico se da al examinar relaciones entre variables, a través de diagramas de dispersión *(scatter plots):*

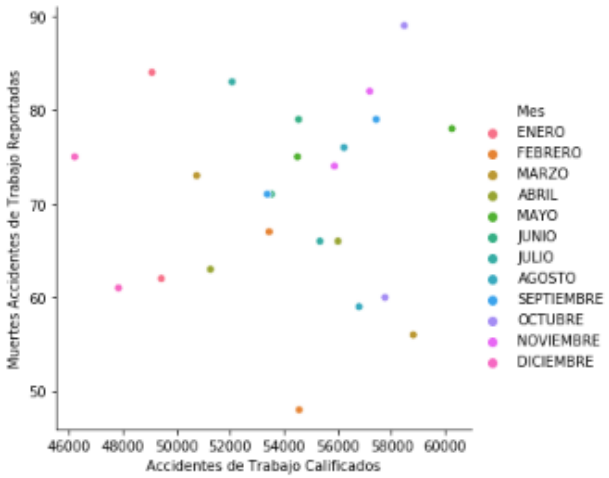


Fig. 5. Incluso cruzando 2 variables numéricas y una categórica, no es posible vislumbrar relaciones entre variables

Sin embargo, a través de otros diagramas, es posible determinar situaciones atípicas como que alrededor de ¼ de las incidencias de muerte de enfermedades laborales, ocurrió en 1/12 parte de los meses:

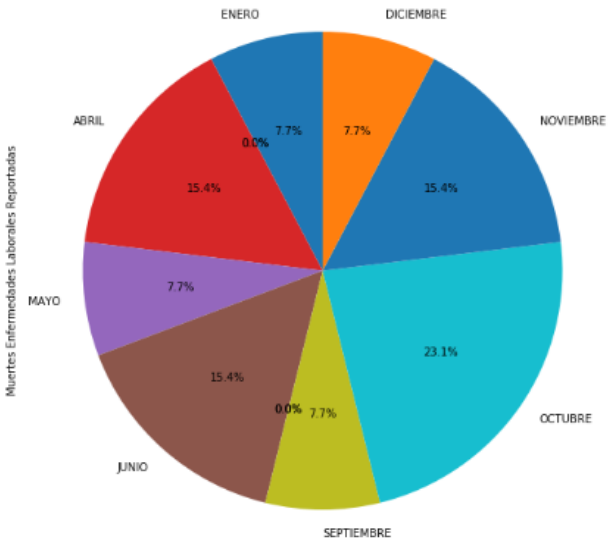


Fig. 5. La zona azulada clara muestra una proporción del 23,1 % sobre el total de muertes por enfermedades laborales

A manera de conclusión para este ejercicio se tiene pues una tendencia incremental hacia el 2018, una proporciona atípica de muertes por enfermedad laboral, en octubre, pero que a su vez dichas consideraciones deben tener en cuenta la altísima variabilidad en los atributos para afiliados, hecho que lo convierte en un estimador con poca confiabilidad

## II. DATASET ROBUSTO

### A. *Etapa de Revisión*

Al igual que los lineamientos seguidos para el tratamiento de datos del conjunto de inicial, es necesario llevar a un proceso de evaluación de consistencia de la información, pero ahora referentees al dataset ligado al *endpoint* https://www.datos.gov.co/Salud-y-Protecci-n-Social/Estadisticas-Riesgos-Laborales-Positiva/kwqa-xugj, para este se presenta a diferencia del primero una mayor concurrencia de observaciones (>30000 filas), sumado a la poca documentación respecto a las variables presentes en el dataset, son elementos que dificultan exponencialmente la búsqueda de relaciones fehacientes entre variables:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| CODIGO_DE_LA_ARL | 44750.0 | 1.423000e+03 | 0.000000e+00 | 1423.0 | 1423.0 | 1423.0 | 1423.0 | 1423.0 |
| AÑO_DE_INFORME | 44750.0 | 2.019000e+03 | 0.000000e+00 | 2019.0 | 2019.0 | 2019.0 | 2019.0 | 2019.0 |
| MES_DE_INFORME | 44750.0 | 1.100000e+01 | 0.000000e+00 | 11.0 | 11.0 | 11.0 | 11.0 | 11.0 |
| ACTIVEC | 44750.0 | 2.873962e+06 | 1.322368e+06 | 1014001.0 | 1751201.0 | 2521902.0 | 3742102.0 | 5930901.0 |
| RELA_DEP | 44750.0 | 4.844212e+01 | 6.093606e+02 | 0.0 | 0.0 | 4.0 | 15.0 | 80686.0 |
| RELA_INDEP | 44750.0 | 8.018503e+00 | 3.616710e+02 | 0.0 | 0.0 | 0.0 | 0.0 | 64195.0 |
| PRESUACCIDETRASUCE | 44750.0 | 2.414078e-01 | 2.900871e+00 | 0.0 | 0.0 | 0.0 | 0.0 | 234.0 |
| MUERTES_REPOR_AT | 44750.0 | 1.340782e-04 | 1.157857e-02 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| NUEVAPENSIOINVA_R_AT | 44750.0 | 1.340782e-04 | 1.157857e-02 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| NUEVAPENSIOINVA_R_EL | 44750.0 | 0.000000e+00 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| INCAPERMAPARCIAR_AT | 44750.0 | 3.687151e-03 | 7.480286e-02 | 0.0 | 0.0 | 0.0 | 0.0 | 5.0 |
| INCAPERMAPARCIAR_EL | 44750.0 | 2.659218e-03 | 9.199126e-02 | 0.0 | 0.0 | 0.0 | 0.0 | 12.0 |

Fig. 6. La primera columna marca un conteo de 44750 registros

En búsqueda de información relevante, se obtiene un listado con los departamentos donde han ocurrido reporte de muertes en accidentes de trabajo, obteniendo:

```
Int64Index([0, 1], dtype='int64') [592   1] Bogotá, D.C.
Int64Index([0, 1], dtype='int64') [1025    1] Bolívar
Int64Index([0, 1], dtype='int64') [145   1] Guaviare
Int64Index([0, 1], dtype='int64') [3105    1] Boyacá
Int64Index([0, 1], dtype='int64') [1356    1] Nariño
Int64Index([0, 1], dtype='int64') [3476    1] Valle del Cauca
```

Fig. 6. Las ciudades mostradas al final de la tala, son los departamentos han habido ocurrencias de muertes en accidentes de trabajo, no hay más de uno para ningún departamento, y un total de 6 accidentes fatídicos

Para incapacidad permanente y parcial (como columna de la tabla), se tiene una concurrencia mayor en Bogotá, donde se presenta las siguientes categorías no documentadas:

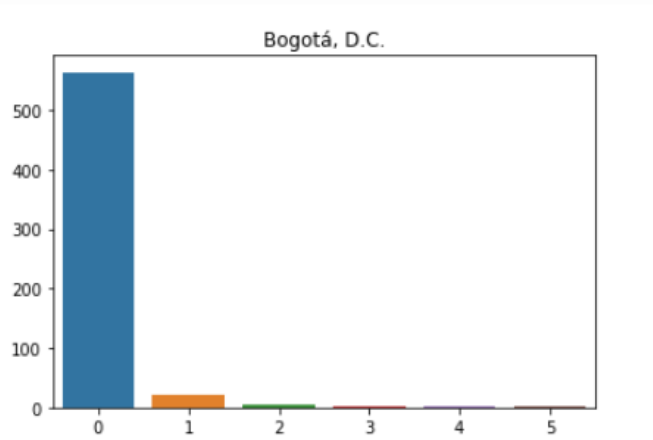


Fig. 7. Para incapacidad permanente y/o parcial, se cuantifica ninguna (0) y otras desconocidas, que no se cuentan documentadas, situaciones de este tipo restringen el alcance de un análisis de estas características

## III. APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING

Para ello se toma como matriz X de características, las variables numéricas de la tabla a excepción de target (ojetivo), MUERTES_REPOR_AT, PRESUACCIDETRASUCE, depedencia e idependencia y otras categorías referentes a la observación como código de ARL y fecha del informe, para dicha categoría:

```
In [11]: Feature = df_t[['NUEVAPENSIOINVA_R_AT','NUEVAPENSIOINVA_R_EL','INCAPERMAPARCIAR_AT','INCAPERMAPARCIAR_EL']]
         y = df_t['MUERTES_REPOR_AT'].values
```

Fig. 8. Se encuentran las variables tomadas en cuenta para X, y

Seguidamente es necesario realizar un balance de la clase objetivo (y) , con el objetivo de garantizar que el algoritmo aprenda forma homogénea de ambas clases, para hacerlo, se importa SMOTE de imblearn.over, para sobresamplear ambos conjuntos pares(X,y):

```
0    44744
1        6
Name: MUERTES_REPOR_AT, dtype: int64

1    44744
0    44744
```

```
Name: MUERTES_REPOR_AT, dtype: int64
```

Fig. 9. Se muestra como se sobresamplea los valores de la variable, hasta que sean iguales las cantidades en la etiqueta binaria

Cuando la data esta preparada, ya es posible instanciar el modelo de IA en cuestión, para este caso en particular, del segundo dataset,:

**Arbol de deciones**

Este modelo clásico de teoría de decisiones, baja su entropía con el nivel de información aprendida, sobre lacaracterística con mayor ganancia de información, de esta forma, el criterio de entropía está definido por:

Entropy = -p(B)*log(P(B))-p(a)(log(p(A))

Bajo este criterio el modelo aprende y clasifica, arrojando los siguientes resultados:

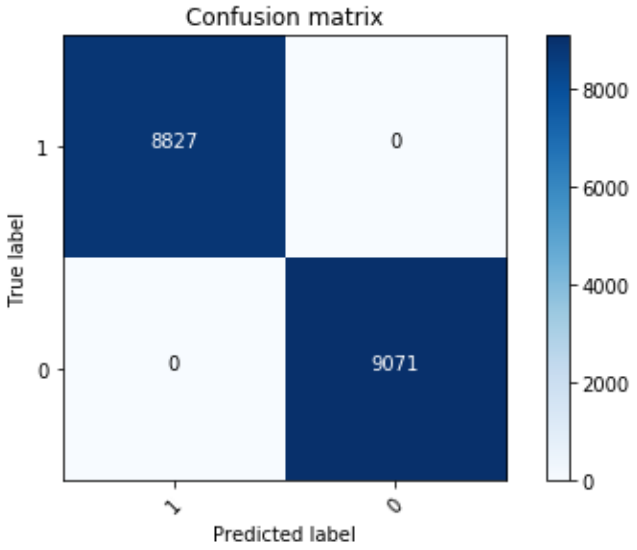


Fig. 9. La matriz de confusión busca precisamente tener una diagonal en cero y la otra ennúmeros, este resultado en particular indica que el modelo predijo correctamente todas las clases del vector y_test, equivalente a una recall y precisión de 1

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 9071 |
| 1 | 1.00 | 1.00 | 1.00 | 8827 |
| accuracy | | | 1.00 | 17898 |
| macro avg | 1.00 | 1.00 | 1.00 | 17898 |
| weighted avg | 1.00 | 1.00 | 1.00 | 17898 |

Fig. 20. Otras medidas de desempeño como la exactitud y precisión se calculan directamente de la suma de fila y columnas entre el elemento que se quiere considerar, una vez más, este resultado indica un precisión y exactitud del 100%

**Support Vector Machine (SVM)**

El empleo de la técnica del SVM supone un mapeo a nivel dimensional más alto de las características (atributos) para encontrar separaciones entre observaciones, incluso si estas no se encuentran linealmente separadas, por tanto, el mecanismo de cálculo se basa en encontrar el hiperplano que mejor pueda separar las observaciones en categorías distintas, es especialmente útil para tareas de clasificación de etiqueta binaria:

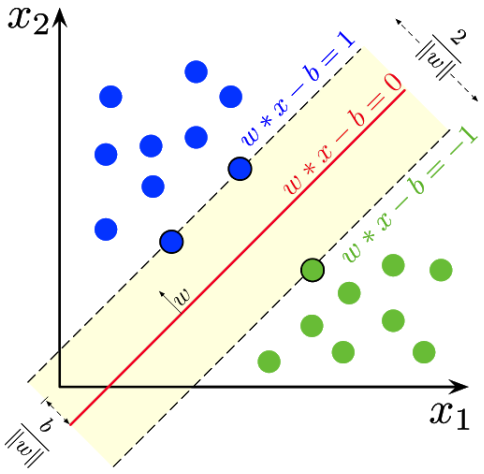


Fig. 9. Principio de operación de SVM, dado el hiperplano w*x-=0 con gradiente w

Empleando este approaching se obtienen los siguientes resultados:

```
from sklearn import svm
kernels_name = ['linear','poly','rbf','sigmoid']
y_hat_2 =[]
results_2 = []
for j in kernels_name:
    y_hat_2 = svm.SVC(kernel = j).fit(X,y).predict(X_test)
    print (classification_report(y_test, y_hat_2))
```

Fig. 21. Aplicación del método con kenels diferentes, confluyen todas en un recall y precisión del 100%, sigmoid function es también trasversal a otro método de clasificación, como es la regresión logística

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 9071 |
| 1 | 1.00 | 1.00 | 1.00 | 8827 |
| accuracy | | | 1.00 | 17898 |
| macro avg | 1.00 | 1.00 | 1.00 | 17898 |
| weighted avg | 1.00 | 1.00 | 1.00 | 17898 |

Fig. 22. Siendo una tabla idéntica del numeral 20, este indicador muestra una vez la predicción correcta de todos los casos

**CONCLUSIONES**

Al emplear técncias de análisis gráficas, de aprendizaje automático supervisado y un análisis particular, es posible llegar a una conjunción de resultados, que en primera instancia muestra la necesidad imperante de información actualizada y precisa sobre las variables publicadas, ya que varios aspectos

del análisis depende de la claridad de este tipo de información, es por ello, que si bien un modelo de inteligencia artifical simple puede clasificar correctamente el 100% de las veces, el desconocimiento explícito de la naturaleza de las variables, no nos permite contextualizar este resultado lo suficiente para evaluar su impacto, para manera de finalización se da predicciones echas por la IA y_hat y las reales y_test:

```
print(yhat[0:5]) #some predictions
print(y_test[0:5])
```

```
[0 0 0 1 1]
[0 0 0 1 1]
```

Fig. 23. Muestra de predicciones, como es notable, la predicción es la misma que el valor real y_test