Entrega 2

JUAN PABLO ARENAS VÉLEZ

Tutor: RAUL RAMOS POLLAN

Introducción Inteligencia Artificial 2022-1

Universidad de Antioquia

2022

Contenido

1. Planteamiento del problema

Conocer las causas de muerte de la humanidad es de vital importancia, ya que esto nos puede permitir estudiar cuales son las principales afecciones que se pueden presentar y por qué no, prevenir por medio de tratamientos o prácticas sanas. Pero para llegar a ello se necesitan recursos de investigación, financiación de proyectos de salud preventiva y predictiva. Para ello se pretende analizar cuál causa de muerte puede ser potencialmente más peligrosa, y con ello desarrollar un modelo que permita mejorar las estimaciones para así generar interés en dichos desarrollos e investigaciones que permitan ejecutar acciones que mejoren la calidad de vida de las personas y mitiguen esas estadísticas de muerte.

2. Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos del número anual de muertes por diferentes causas alrededor del mundo. El dataset está compuesto por un conjunto de archivos .csv que proporcionan la información requerida, El archivo que contiene los datos del edificio es nombrado cómo:
annual-number-of-deaths-by-cause.csv y contiene la siguiente información principal:

https://www.kaggle.com/datasets/programmerrdai/cancer?select=annual-number-of-deaths-by-cause.csv

Entity: País,

Code,Código del país

Year, Año,

Number of executions (Amnesty International), Número de ejecuciones (Amnistía Internacional),

Deaths - Meningitis - Sex: Both - Age: All Ages (Number), Muertes - Meningitis - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Alzheimer's disease and other dementias - Sex: Both - Age: All Ages (Number), Muertes - Enfermedad de Alzheimer y otras demencias - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Parkinson's disease - Sex: Both - Age: All Ages (Number), Muertes - Enfermedad de Parkinson - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Nutritional deficiencies - Sex: Both - Age: All Ages (Number), Muertes - Deficiencias nutricionales - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Malaria - Sex: Both - Age: All Ages (Number), Muertes - Malaria - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Drowning - Sex: Both - Age: All Ages (Number), Muertes - Ahogamiento - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Number), Muertes - Violencia interpersonal - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Maternal disorders - Sex: Both - Age: All Ages (Number), Muertes - Trastornos maternos - Sexo: Ambos - Edad: Todas las edades ,

Deaths - HIV/AIDS - Sex: Both - Age: All Ages (Number), Muertes - VIH/SIDA - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number), Muertes - Trastornos por consumo de drogas - Sexo: Ambos - Edad: Todas las edades

Deaths - Tuberculosis - Sex: Both - Age: All Ages (Number), Muertes - Tuberculosis - Sexo: Ambos - Edad: Todas las edades),

Deaths - Cardiovascular diseases - Sex: Both - Age: All Ages (Number), Muertes - Enfermedades cardiovasculares - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Lower respiratory infections - Sex: Both - Age: All Ages (Number), Muertes - Infecciones de las vías respiratorias bajas - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Neonatal disorders - Sex: Both - Age: All Ages (Number), Muertes - Trastornos neonatales - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number), Muertes - Trastorno por consumo de alcohol - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Self-harm - Sex: Both - Age: All Ages (Number), Muertes - Autolesiones - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Exposure to forces of nature - Sex: Both - Age: All Ages (Number), Muertes - Exposición a las fuerzas de la naturaleza - Sexo: Ambos - Edad: Todas las edades),

Deaths - Diarrheal diseases - Sex: Both - Age: All Ages (Number), Muertes - Enfermedades diarreicas - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Environmental heat and cold exposure - Sex: Both - Age: All Ages (Number), Muertes - Exposición al calor y al frío ambiental - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Neoplasms - Sex: Both - Age: All Ages (Number), Muertes - Neoplasias - Sexo: Ambos - Edad: Todas Edades ,

Deaths - Conflict and terrorism - Sex: Both - Age: All Ages (Number), Muertes - Conflicto y terrorismo - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Diabetes mellitus - Sex: Both - Age: All Ages (Number), Muertes - Diabetes mellitus - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Chronic kidney disease - Sex: Both - Age: All Ages (Number), Muertes - Enfermedad renal crónica - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Poisonings - Sex: Both - Age: All Ages (Number), Muertes - Intoxicaciones - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Protein-energy malnutrition - Sex: Both - Age: All Ages (Number),Terrorism (deaths), Muertes - Desnutrición proteico-energética - Sexo: Ambos - Edad: Todas las edades ,Terrorismo (muertes),

Deaths - Road injuries - Sex: Both - Age: All Ages (Number), Muertes - Lesiones viales - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Chronic respiratory diseases - Sex: Both - Age: All Ages (Number), Muertes - Enfermedades respiratorias crónicas - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Cirrhosis and other chronic liver diseases - Sex: Both - Age: All Ages (Number), Muertes - Cirrosis y otras enfermedades crónicas del hígado - Sexo: Ambos - Edad: Todas las edades ,

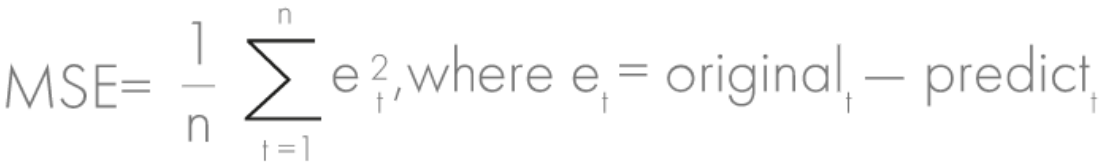
Deaths - Digestive diseases - Sex: Both - Age: All Ages (Number), Muertes - Enfermedades digestivas - Sexo: Ambos - Edad: Todas las edades ,

Deaths - Fire, heat, and hot substances - Sex: Both - Age: All Ages (Number), Muertes - Fuego, calor y sustancias calientes - Sexo: Ambos - Edad: Todas las edades

Deaths - Acute hepatitis - Sex: Both - Age: All Ages (Number), Muertes - Hepatitis aguda - Sexo: Ambos - Edad: Todas las edades

3. Métricas

La métrica con la cual evaluaremos principalmente a nuestro modelo será el ECM (Error cuadrático medio), el cual calcula el valor medio de la diferencia al cuadrado entre el valor real y el predicho para todos los puntos de datos



en cuanto a la métrica de negocio queremos determinar si las predicciones son lo suficientemente confiables para conocer las causas de muerte más recurrentes en los diferentes países y ver si la inversión en la prevención de estas causas vale la pena, y no solo sean causas que no puedan remediarse, se busca llegar al análisis de esta viabilidad, si vale la pena o no invertir en la prevención.

4. Desempeño

El criterio más importante sería predecir que una de las más altas probabilidades de muerte esté relacionada con causas que puedan prevenirse, para garantizar la inversión en un proyecto de prevención. El ideal es que las causas de muerte prevenibles estén por encima del 70%.

5. Preprocesado de los datos

- **Remoción de las lecturas cero de la variable objetivo**

Como se mencionó anteriormente al analizar la variable objetivo, se observó que existen muchas lecturas de valor cero para la variable objetivo (Deaths - Meningitis - Sex: Both - Age: All Ages (Number)). Debido a esto se opta por eliminar las filas que contengan un valor cero para la variable objetivo ya que esto indica que no se tiene una disponibilidad de la medición y estos no podrían utilizarse para entrenar.

```
death_cancer_zero = list(cancer_data[cancer_data['Deaths - Meningitis - Sex: Both - Age: All Ages (Number)'] == 0].index)
cancer_data.drop(death_cancer_zero, axis = 0, inplace = True) #eliminar filas con lecturas cero en la variable objetivo
print('Nuevo tamaño de los datos: ',cancer_data.shape)
```

- **Eliminación de las columnas con muchos datos faltantes**

Del análisis exploratorio de las variables, se encontró que existen variables que contienen gran cantidad de datos faltantes. En este caso se optará por eliminar las columnas en las que los datos faltantes representen el 50% o más de la totalidad de los datos.

```
criterio = len(cancer_data) * 0.5 #criterio para eliminar la 
cancer_data.dropna(axis=1, thresh = criterio, inplace = True)
print('New Shape of cancer_data Data:',cancer_data.shape)
```

- **Relleno de datos faltantes**

Anteriormente se optó por eliminar las columnas que poseían el 50% o más en datos en datos faltantes, sin embargo, aun quedan variables con datos faltantes en el dataset. Para estas columnas se rellenaran los valores faltantes al reemplazarlos por la media de los valores con los que se cuenta para cada columna, primero conoceremos quienes son:

- **Transformación de las variables categóricas**

Las variables categóricas, como lo son en este caso "Entity" pueden ser útiles a la hora de realizar el análisis, sin embargo, no pueden ser usadas en la forma categórica, por lo que se deben convertir en variables numéricas que si podamos utilizar para entrenar un modelo.

```
var_categoricas = ['Entity']
encoder = preprocessing.LabelEncoder()

for i in var_categoricas:

    cancer_data[i] = encoder.fit_transform(cancer_data[i])

print (cancer_data.info())
```

6. Bibliografía

| Kaggle. (2021). Retrieved 16 December 2021, from https://www.kaggle.com/datasets/programmerrdai/cancer?select=annual-number-of-deaths-by-cause.csv

https://fayrix.com/machine-learning-metrics_es