Universidad del Valle de Guatemala
Data Science 1 - Sección 10
Christopher Kevin Sandoval García 13660
María Fernanda Estrada Cornejo 14198
Rodrigo Samayoa Morales 17332
David Uriel Soto Alvarez 17551
Ana Villela 18903
Guatemala 10 de septiembre de 2020

Análisis exploratorio del proyecto 2

Situación problemática y problema científico

- Situación problemática

El estudio de la edad ósea permite a los médicos saber la madurez del sistema esquelético de un niño. La edad ósea se mide en años y meses. El método más utilizado es mediante la radiografía de la mano izquierda, que va desde la muñeca hasta los dedos. Esta imagen es comparada con un atlas estándar del desarrollo óseo normal de niños de la misma edad y sexo. Este estudio normalmente lo solicita un pediatra para evaluar qué tan rápida o lenta es la maduración esquelética del niño, con el fin de determinar si padece de una enfermedad. El problema que tiene este método de comparación es que se pueden dar errores de interpretación, ya que depende absolutamente de la persona que compara. Además, si el niño se encuentra en la etapa temprana de una enfermedad ósea, esta comparación es muy importante.

- Problema científico

El problema científico que se nos presenta es cómo poder minimizar la cantidad de errores que se dan por el método de comparación, que si bien ya se mencionó, se basa en la interpretación de la persona o especialista que está aplicando el método. Por lo tanto, se pretende buscar una manera en la cual se pueda aplicar la tecnología y Data Science con el fin de automatizar este proceso y llegar a generar un modelo que permita hacer una clasificación confiable que permita poder dar diagnósticos más confiables, precisos y exactos para los pacientes. Sin embargo, hay que lograr hacer que el modelo pueda ser confiable ante la problemática mencionada para no tener problemas con respecto a clasificaciones erróneas.

Objetivos

- General

El objetivo de este proyecto es poder realizar un análisis de imágenes de estructuras óseas, en particular del brazo, para poder automatizar este proceso que permite determinar la Edad Ósea (EO) del individuo.

- Específicos

1. Generar modelos y predicciones sobre la Edad Ósea con el uso de Data Science, a partir de técnicas de filtrado y procesamiento de imágenes.

2. Minimizar la cantidad de errores al realizar un diagnóstico de Edad Ósea, para que no se tenga que hacer un análisis subjetivo por parte de especialistas, el cual podría estar basado muchas veces en el sesgo de la persona.

Descripción de los datos

El set de datos a analizar consiste en 12800 imágenes de radiografías de mano, cada una acompañada de su id, género y edad ósea en meses.

- Variables

Nombre	Categoría	Descripción	Ejemplo
id	Categórica nominal	Indica a qué imagen están relacionados los datos de género y edad ósea. Solamente es un identificador.	rango: 1377 - 15.6k 1377, 1378, 1379...
boneage	Numérica discreta	Indica la edad ósea en meses de la imagen.	rango: 1 - 288 180, 12, 94, 120...
male	Categórica binaria	Indica si la radiografía es de sexo masculino o no. Es un valor booleano; si es True es masculino, si es False es femenino.	True, False

- Operaciones de limpieza

El set de datos es principalmente numérico, por lo que no se necesitaron muchas operaciones de limpieza. Sin embargo, para simplificar aún más, se cambiaron los valores de la columna male. Se realizó lo siguiente:

1. Verificar que solo existan datos numéricos en las columnas de id y boneage.
2. Verificar que solo existan valores de True y False en la columna de male.
3. Cambiar los valores de True y False a 1 y 0, respectivamente en la columna male.
4. Verificar que hay la misma cantidad de imágenes y datos relacionados.

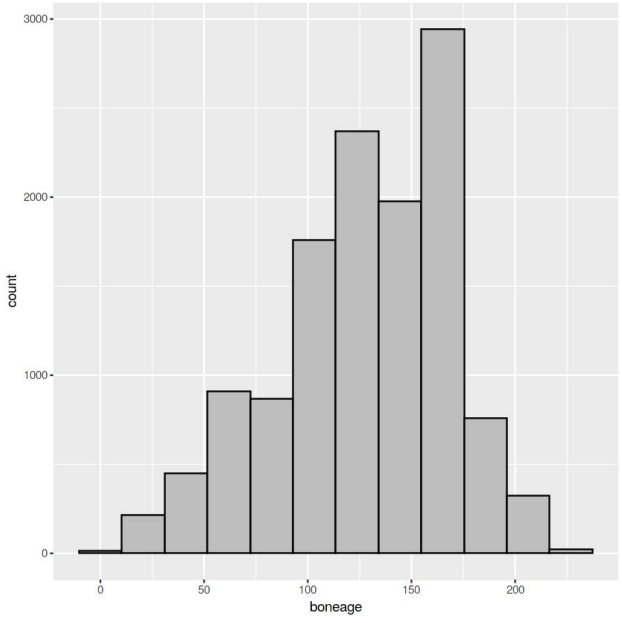
Análisis exploratorio inicial

- Variable cuantitativa: boneage
    - Resumen

```
    boneage
Min.   :  1.0
1st Qu.: 96.0
Median :132.0
Mean   :127.3
3rd Qu.:156.0
Max.   :228.0
```

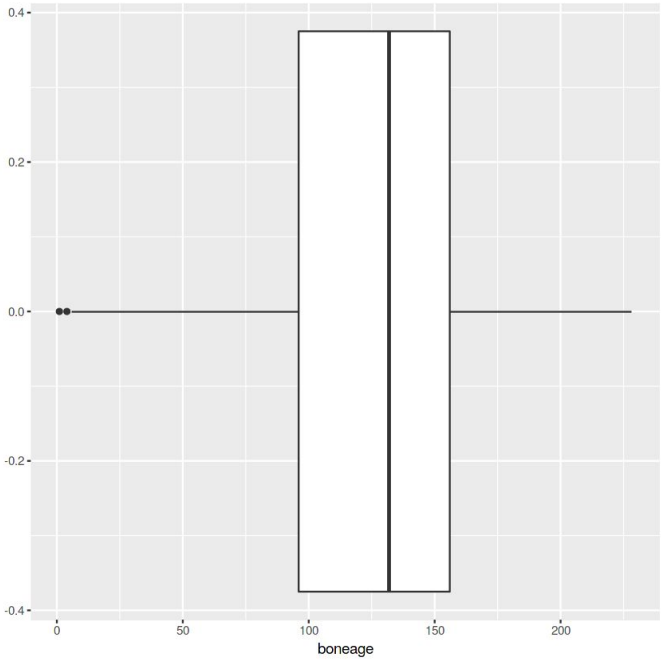
Como se puede observar en la tabla de resumen que se realizó para la variables “bonage”, se puede ver que esta tiene un rango de valores entre 1 a 228, lo cual vendría siendo la edad ósea en años en que está acotada el dataset. También es posible observar que el primer cuartil de los datos está entre el dato en los 96 años, que el segundo cuartil o la mediana está a los 132 años, y por último el tercer cuartil está a los 156 años. Esto podría darnos una idea de que el primer cuartil o las edades óseas más bajas son las que tienen mayor presencia o variedad en este dataset. Por otro lado es interesante ver que la media, que es 127 años es muy similar a la media por lo que tal vez podríamos tener una distribución normal de la muestra de datos, o quizás existan algunos datos atípicos que generan este resumen estadístico.

    - Histograma



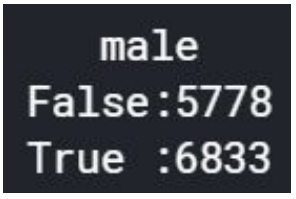
Se puede observar una distribución casi normal en los datos de la variable boneage, con media en 127 años, tal y como se observó en el resumen de los datos. No es completamente normal debido a que existe un pequeño sesgo a la derecha, pero aun así este sesgo es parte de los datos centrales, por lo cual se podría decir que los datos estan los bastante normalizados. Cabe destacar que las edades “extremos” son poco frecuentes y que mientras más nos acercamos a la media, más casos o datos se tienen.

- Diagrama de caja y bigotes



En esta gráfica podemos observar que no hay muchos datos atípicos en la variable de boneage y que esta variable tiene un rango intercuartil aproximadamente entre 100 y 150. También podemos observar que hay un ligero sesgo a la derecha en los datos el cual es normal dentro de lo que se puede esperar en un dataset que tiene una distribución casi normal.

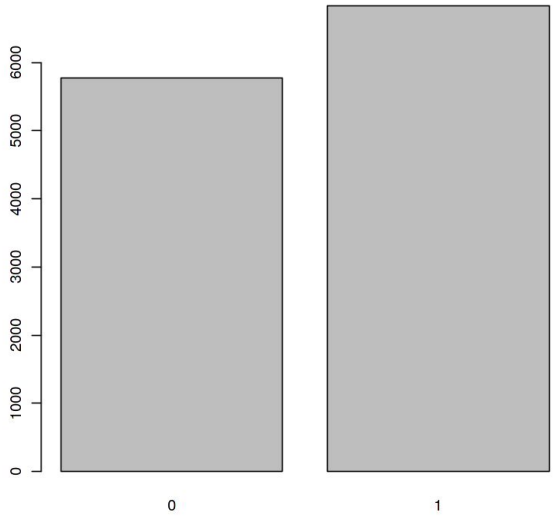
- Variable cualitativa: male
  - Tabla de frecuencia



Se puede observar en la tabla de frecuencia de la variable “male”, que existen únicamente dos posibles valores: True y False. True indica que la radiografía pertenece a un hombre y False indica que pertenece a una mujer. Como indica la tabla, en el set de datos hay más muestras de hombres que de mujeres, con una diferencia de 1055 datos. A pesar de que es una diferencia amplia en los datos, dado que se está trabajando en el orden de los miles, cabe dentro de lo normal tener una diferencia de este tipo, y que los datos sean útiles para el estudio.

Dado que el texto que contiene la variable es poco útil y es una categoría binomial, entonces se decidió seguir analizando esta variable haciendo la sustitución de True y False por 1 y 0 respectivamente.

- Gráfico de barra - Donde 0 es False y 1 es True



De una forma más gráfica, se muestra que hay más cantidad de hombres (variable 1) que de mujeres (variable 0). Sin embargo, considerando la cantidad de datos de cada uno, la diferencia es mínima. Por lo mismo que se mencionó anteriormente, se determinó que el conjunto de datos está lo bastante normalizado y distribuido para realizar un estudio en donde se pueda considerar tanto a hombre como mujeres.

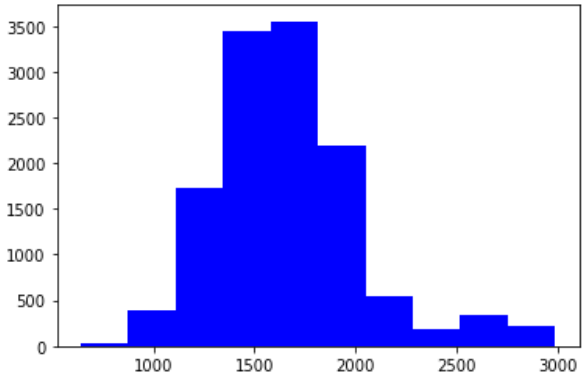
- Tabla de proporción - Donde 0 es False y 1 es True

```
        0         1
0.4581714 0.5418286
```

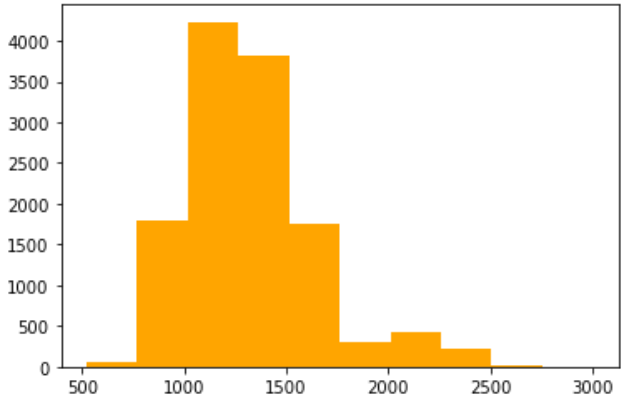
En esta tabla de proporción es más notorio que la diferencia entre la cantidad de datos para hombres y mujeres es mínima, ya que la proporción de ambos es cercana al 50%. Con esto en mente, podríamos inferir que la modelación que se hará a partir de estos datos para poder clasificar y dar diagnósticos de edad ósea, será más probable que no se tengan sesgos causados por el sexo de la persona, ya que ambos sexos están siendo tomados en cuenta en propociones casi iguales.

- Imágenes de radiografías:

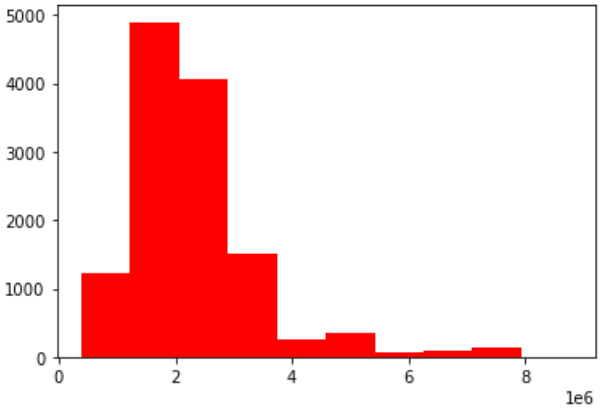
  - Histogramas



Este histograma muestra el rango de valores de height del set de imagenes proporcionado. La altura más grande es de 2989 pixeles y la más pequeña de 640 pixeles. También se puede observar que la mayoría de imágenes tienen altura entre 1200-2000 pixeles.



Este histograma muestra el rango de valores de width del set de imagenes proporcionado. La anchura más grande es de 3001 pixeles y la más pequeña de 521 píxeles. También se puede observar que la mayoría de imágenes tienen anchura entre 800-1600 pixeles.



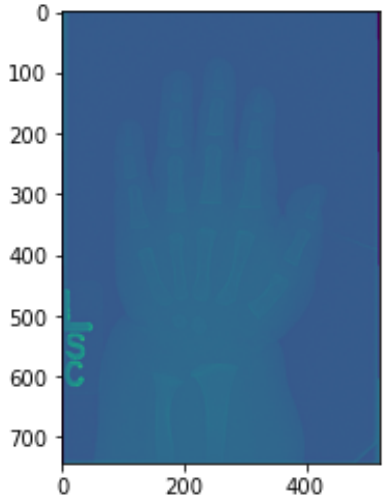
Este histograma muestra el rango de valores de píxeles del set de imagenes proporcionado. La cantidad de píxeles más grande es de 8765921 y la más pequeña de 387103. También se puede observar que la mayoría de imágenes tienen cantidad de pixeles entre 0.5-3.5 millones.

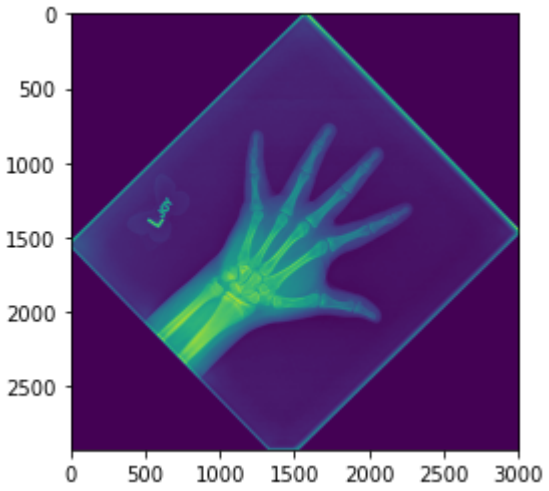
- Resumen

  A fín de tener recopilados los datos anteriormente mencionados, se generó un resumen como sigue

```
La altura mas grande es:  2989
La altura mas pequeña es:  640
La anchura mas grande es:  3001
La anchura mas pequeña es:  521
La cantidad de pixeles mas grande en un imagen es:  8765921
La cantidad de pixeles mas pequeña en un imagen es:  387103
```

- Ejemplos de radiografías
  Ya con los datos de cuál es la altura y anchura más grande y más pequeña, se procedió a identificar la diferencia entre ambas imágenes. Se observa que la radiografía más pequeña está muy borrosa y no se distinguen bien los huesos de la mano. A diferencia de esta, la radiografía más grande muestra gran detalle de los huesos. Es importante notar que la posición de las manos en ambos casos es totalmente distinta, por lo que se debe considerar rotar y recortar ciertas imágenes. A continuación se muestran ambas radiografías.





Hallazgos y conclusiones

En cuanto a la variable cuantitativa “boneage” se determinó que los datos que contiene el dataset a utilizar están distribuidos de una manera casi normal. Esto nos ayudará a que el modelo de clasificación y diagnóstico que se genere a partir de los datos pueda ser más preciso y objetivo. Esto debido a que el modelo tendrá en cuenta que datos de edades óseas son más comunes y cuales son los que menos presencia tienen.

A partir de lo anterior, también cabe destacar que tanto la media como la mediana de la edad ósea son muy cercanas, lo cual también dá una idea sobre cómo el modelo podrá tener un mejor desempeño al conocer que los datos que más frecuencia tienen son los mismo que se concentran en la mediana y por ende la media. Es por esto que al momento en que se haga el procesamiento de imágenes se podrá distinguir y clasificar de mejor manera los casos para diagnosticar. Y es importante mencionar que al analizar el gráfico de caja y bigotes, no se nota que haya muchos datos atípicos, lo cual podrá acotar de mejor manera los rangos de edades óseas que se presentan con mayor frecuencia y saber distinguir entre un caso atípico, aunque no se tenga mucha información de estos casos.

Por otro lado, en la variable cualitativa “male” se observó principalmente que la proporción de hombres y mujeres en el set de datos es casi igual, por lo que no habrá sesgo en esta variable. Sin embargo, aunque la proporción es similar, sí existe una diferencia en la cantidad de datos al haber más hombres que mujeres en el set.

Esta pequeña diferencia podría hacer que el modelo tenga fallos y la exactitud del mismo baje, pero podría ser lo esperado, a modo de no tener un modelo que cause overfit, y que el modelo pueda ser lo suficientemente capaz de poder clasificar y dar diagnósticos de edad ósea, sin caer a errores críticos, cómo por ejemplo dar una edad ósea baja cuando en realidad es alta, por ejemplo.

Por lo tanto, se concluye que los siguientes pasos a seguir son:

1. Realizar el procesamiento de imágenes (cambiar de tamaño, aplicar rotaciones, filtros, entre otros) con el fin de poder transformar y estandarizar las mismas, para que puedan ser procesadas y entendidas de mejor manera por alguno de los modelos de clasificación y predicción estudiados.
2. Determinar qué modelo se acopla mejor para lograr clasificar y predecir la edad ósea a partir de la dataset que se limpió en esta fase.
3. Entrenar el modelo con el set de datos de training.
4. Predecir la edad ósea del set de datos de test.
5. Realizar un análisis de resultados con el fin de determinar qué tan preciso es el modelo para clasificar y dar diagnósticos con respecto a la edad ósea de una persona.
6. Se esperaría que los resultados fueran lo más exactos y precisos posibles, y que factores como el sexo no influyeran en lo bueno o malo que fuera el modelo, ya que en el análsis exploratorio se determinó que para la muestra de datos esto era un factor bastante balanceado en el dataset.

Referencias

1. Durani, Y. s.f. *Radiografía: estudio de la edad ósea*. Consultado el 07/09/2020 de https://kidshealth.org/es/parents/xray-bone-age-esp.html#:~:text=Qu%C3%A9%20es,una%20peque%C3%B1a%20cantidad%20de%20radiaci%C3%B3n.
2. Navarro, M.; Tejedor, B.; López, J. 2014. *El uso de la edad ósea en la práctica* clínica. Consultado el 07/09/2020 de https://www.elsevier.es/es-revista-anales-pediatria-continuada-51-articulo-el-uso-edad-osea-practica-S1696281814702045
3. Pérez, R. 2011. *Valoración y utilidad de la edad ósea en la práctica clínica*. Consultado el 07/09/2020 de https://fapap.es/articulo/180/valoracion-y-utilidad-de-la-edad-osea-en-la-practica-clinica