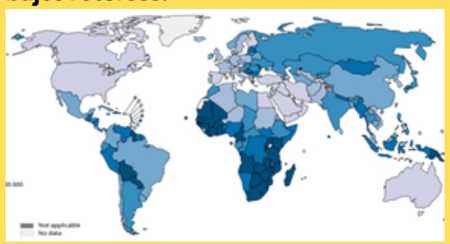
Diagnostico de cáncer de cuello uterino

SUÁREZ MARTÍNEZ M.A.

Introducción

El cáncer de cuello uterino figura entre los principales cánceres ginecológicos del mundo. Según los datos actuales de la OMS, el cáncer de cuello uterino es el segundo cáncer más frecuente en mujeres con un estimado de 570.000 casos nuevos en 2018, lo que representa el 14,7% de todos los cánceres femeninos, es la 4° neoplasia más frecuente en mujeres a nivel mundial, las tasas de incidencia más altas se producen en América Central y del Sur, en el África subsahariana y en el sudeste asiático .

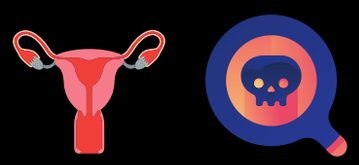
En un estudio reciente que incluyó 38 países de los 5 continentes se mostró una sustancial disminución de la tasa de incidencia y mortalidad en los países con mayor tasa de ingresos, mientras estas tasas se estabilizaron o incluso aumentaron en aquellos países de bajos recursos.



Países con mas casos

Objetivos

Buscamos servir como apoyo en los hospitales y en cualquier servicio medico para ayudar en el diagnostico temprano de esta enfermedad, tomando en cuenta las características del paciente queremos saber cuales son mas propensos a padecer esto, crearemos un algoritmo que prediga el resultado de la biopsia y a su vez queremos saber características de los pacientes que resultan padecer esta enfermedad.



Resultados

Buscando las características de las pacientes que resultaron tener esta enfermedad apoyándonos con las graficas de la derecha notamos que:

- Las pacientes suelen tener un promedio 28 años, 2 parejas sexuales, 2 embarazos y empezaron su vida sexual al rededor de los 17 años. (Tabla 1)
- Si vienen tienen un promedio de 2 embarazos también hay muchas pacientes con 1 y 3 embarazos, llama la atención que todas las pacientes que dieron positivo en la biopsia tenían por lo menos un embarazo. (Grafico 2)
- Es muy raro detectar pacientes menores a los 20 años, entre los 20 y 40 años las mujeres pueden tener un gran riesgo de sufrir la enfermedad. (Grafico 3)

Tabla (1)

	Edad	Numero de parejas sexuales	Edad primer relacion sexual	Numero de embarazos
media	26.820513	2.509091	17.072727	2.472727
moda	28	2	17	2

Grafico (2)

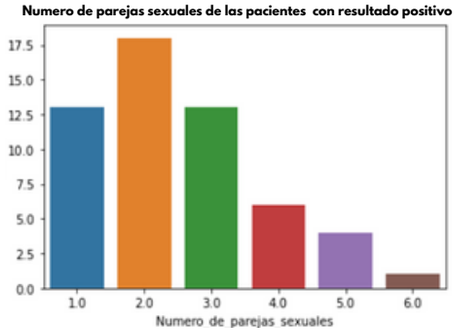


Grafico (3)

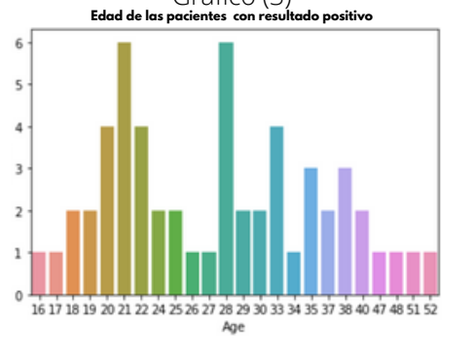
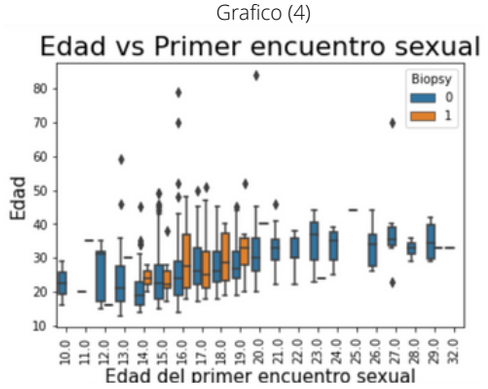


Grafico (4)



En el Grafico (4) notamos que aquellas que tuvieron su primera relación sexual entre los 15 y 20 años de su vida son más propensos a dar positivo en la prueba de biopsia y esas personas se encuentran predominantemente en el grupo de edad de 20 a 35 años.

El objetivo principal era encontrar un modelo que pudiese predecir si una paciente podría o no tener un resultado positivo en la biopsia, implementando la técnica de bosque aleatorio tomando como la variable de salida la biopsia y variable de entrada algunas variables seleccionadas, logramos un algoritmo con 91.6% de precisión y una exactitud de 98.8%

	Modelo	Exactitud	Puntaje F1	Sensibilidad	Precision
0	Arbol de decision (Datos balanceados)	0.965116	0.790698	0.894737	0.708333
1	Bosque Aleatorio (Datos balanceados)	0.965116	0.780488	0.842105	0.727273
0	Arbol de decision (Variables especificas)	0.980620	0.814815	0.846154	0.785714
1	Bosque Aleatorio (Variables especificas)	0.988372	0.880000	0.846154	0.916667

Aquí podrán encontrar el algoritmo implementado

Conclusiones

Este algoritmo no busca sustituir las pruebas medicas, mas bien buscamos complementar y ayudar a la detección temprana de esta enfermedad, con este análisis notamos que tener una temprana actividad sexual es un gran factor para desarrollar ese padecimiento a futuro sin embargo rara vez se desarrolla antes de los 20, esto nos dice que se tiene muchos años para que la enfermedad ataque, por lo que se tiene mucho tiempo para actuar contra la enfermedad, los embarazos también aumenta el riesgo, nos damos cuenta de esto porque todas las pacientes con resultado positivo tenían al menos un embarazo.

En la mayoría de los casos esta dolencia es predecible, y gracias al algoritmo podemos ayudar a saber si una paciente tendrá esta enfermedad o no con una exactitud del 98.8%.

El sistema de salud es deficiente y muchas mujeres mueren por la detección tardía de este padecimiento, espero con optimismo que el aprendizaje de maquina ayude a anticipar con un buen diagnostico a todas aquellas que sean mas propensas a padecerlo y a si ayudar a mejorar el sistema de salud.

Recursos

La base de datos empleada en este estudio fue proporcionada por el Hospital Universitario de Caracas en Caracas y subida a Kaggle por el doctor Senthamarai Kannan con el nombre “Cervical Cancer Risk Classification”

Metodologia

La base de datos “Cervical Cancer Risk Classification” cuenta con 857 registros de pacientes y 36 variables con las características de su respectivo paciente de las cuales se eliminaron 3 de la base de datos, 2 debido a que tenia mas del 80% de datos nulos, y una por que no aportaba nada a nuestra base de datos.

Para la limpieza se contabilizo los datos nulos de cada columna, y debido a que no eran muchos datos nulos decidimos rellenarlos con la mediana de su respectiva columna.

Utilizaremos la técnica de bosque aleatorio para predecir si un paciente tiene la enfermedad.