Predicción de tendencia al desarrollo de células cancerosas con inteligencia artificial Juan Carlos Garfias Toyar

A01652138@itesm.mx

Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, México

1. Abstract

Este proyecto propone crear un método para la clasificación de individuos con tendencia a desarrollar células cancerosas a través de una predicción inteligencia artificial. Se utilizó un método de clasificación binaria a partir de un modelo con redes neuronales y *cross validation*. Esto con el fin de crear y entrenar modelos que puedan clasificar con precisión la tendencia de un paciente a padecer cáncer pulmonar.

2. Introducción

El cáncer se ha convertido en una de las primeras causas de muerte a nivel mundial, la Organización Mundial de la Salud estimó un total de 8.2 millones de muertes en donde los principales casos de esta enfermedad se ubicaban en el pulmón, hígado, estomago, colon y mama (OMS, 2019). Existen diferentes causas para que un paciente padezca cáncer. Este se produce por la transformación de células normales en células tumorales. Este proceso sucede durante la replicación del ADN. El cuerpo humano es capaz de regular la cantidad de células, esto debido a que las células producen señales para controlar cuantas debe de haber y cada cuanto se deben de reproducir. En caso de que una de estas señales falle las células comienzan a crecer descontroladamente y a multiplicarse indefinidamente creando un tumor. En otros casos como la leucemia el cáncer inicia en el torrente sanguíneo donde se reproducen o se adhieren a la medula ósea. Posterior a ello el cáncer comienza a gestarse.

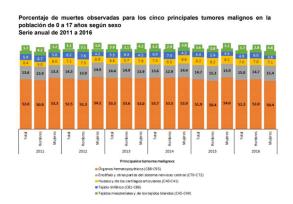
Una mutación es conocida como un cambio que sucede en los genes durante la división celular. Estas mutaciones pueden afectar a la comunicación y por lo tanto al crecimiento. En general debe de haber media docena de diferentes mutaciones para que una célula normal pase a ser cancerosa. La mutación de las células puede ser resumida como una sobre producción de proteínas o una carencia de producción (Research UK, 2019).

Existen diferentes tipos de tumores, los tumores malignos y benignos. Los tumores benignos no son considerados cáncer mientras que los malignos sí. Los benignos crecen en una sola parte del cuerpo y no pueden aparecer ni invadir otras secciones. Sin embargo, si son de un volumen considerado pueden ocasionar daños por presión a órganos vitales (NIH, 2019). Por otro lado, los malignos pueden ocasionar que exista metástasis, es decir que exista una diseminación del cáncer, donde las células cancerosas se separan de este tumor primario y viajan a través del sistema sanguíneo o linfático para formar otros tumores en nuevos órganos (NIH, 2019).

El presente documento pretende exponer métodos de detección temprana antes de que la metástasis e inclusive la gestión de un tumor comience a través del análisis de condiciones e indicadores de salud de un paciente. Esto a través de una técnica computacional basada en redes neuronales, las cuales imitan el procesamiento de información para encontrar patrones en datos, indicando si tiene tendencia a padecer la enfermedad.

3. El cáncer en México

El 99% de los pacientes con cáncer de pulmón en México son diagnosticados cuando ya se encuentran en etapas de metástasis. De acuerdo con la Organización Mundial de la Salud (OMS) existen 2.7 millones que padecen cáncer de pulmón. En el caso de México este cáncer es el noveno mas frecuente y causa alrededor de 18 muertes diarias (EFE, 2019). En el periodo de 2011 a 2016 el 50% de las muertes por tumores malignos en la población de 0 a17 años se debía a cáncer hematopoyéticos, de igual modo se mostro una tendencia general por parte de los hombres a tener una mayor tasa de defunción.



Estadísticas de mortalidad (INEGI, 2018)

De igual manera la INEGI (2018) reportó que el cáncer de pulmón fue una de las principales causas de mortalidad en cuanto tumores malignos en la población de 30 a 59

años. Además de esto se reportó una relación directa en cuanto al desarrollo de esta situación medica con el habito de fumar. Además de esto se demostró que el tabaco puede tener efectos en el desarrollo de cáncer de vejiga, riñón, páncreas, entre otros.

3.1 Procedimientos para el diagnóstico de pulmón

Uno de los métodos comunes para el estudio de un tumor pulmonar o anomalía es una biopsia. A pesar de esto existen otros métodos sin intervención quirúrgica efectuados por médicos. El diagnóstico es sospechado por los síntomas y signos causados por efectos locales. Tales como el crecimiento local del tumor, invasión u obstrucción de estructuras vecinas. crecimiento de los ganglios linfáticos y de las metástasis. De igual manera existe la posibilidad de encontrar anomalías a través de radiografías simples de tórax (Sainz B.,2006). En general se busca que se confirme o descarte la sospecha del tumor, determinar su histología, comprobar si existe metástasis y determinar su estadio.

En la actualidad se emplean métodos con índole más tecnología que los tradicionales tal como es el caso de la utilización de imágenes por el método "screening". El cual utiliza escaneo tipo CT de pecho de baja dosis, con la cual es posible a través de un monitoreo y análisis determinar si se esta desarrollando un tumor en el paciente (Chesnutt, Asha N., et al..,2019). Esta técnica requiere de una alta habilidad y conocimientos por parte del medico para poder realizar un diagnostico adecuado por lo que cualquier uso incorrecto del equipo o

alteración de las imágenes puede dar pie a errores en el diagnóstico.

3.2 Relación del alcohol y tabaco con el cáncer pulmonar

El uso de tabaco es el factor de riesgo mas importante para el cáncer de pulmón. Se establece que hasta el 20% de las muertes por cáncer pueden ser prevenidas por la eliminación de esta sustancia (Pisani P, Bray F, Parkin DM ,2002). El riesgo acumulado de padecer cáncer por tabaco es 30% mayor a la de un 1% a aquellos que no la consumen (Mattson ME, Pollack ES, Cullen JW., 1987). El riesgo de cáncer de pulmón es proporcional a la magnitud del consumo de cigarros como también de el numero de cajetillas por día, la edad del paciente y la susceptibilidad y predisposición genética también son roles fundamentales en cuanto al padecimiento de este tipo de cáncer (De la Cruz, Charles S., et al., 2015). La agencia internacional de investigación del cáncer (IARC) identificó un mínimo de 50 químicos cancerígenos. Estos se combinan con el ADN formando aductos y por lo tanto dando pie a mutaciones genéticas (Smith CJ, Perfetti TA, Rumple MA, Rodgman A, Doolittle DJ., 2000).

La edad del paciente es también un papel importante en cuanto a la probabilidad de desarrollar cáncer. Existe un 9.3% para el rango de 65 años en adelante, 21.4% para el rango de 18 a 24 años, 23.7% para el rango de 25 a 44 años, 22.6% para el rango de 45 a 65 años. La expectativa de vida de 5 años por cáncer de pulmón se ve reducida de acuerdo a la edad del paciente (Dube S, McClave A, James C, Caraballo R, Kaufmann R, Pechacek T., 2009).

Existen complicaciones medicas por el alcoholismo tales como cambios en el metabolismo dando como resultado daños en tejidos de órganos, malnutrición, deterioro de células, entre otros. La diuresis asociada con el consumo de bebidas alcohólicas es causada por la inhibición de la liberación de la hormona antidiurética desde la hipófisis posterior. El alcohol también puede aumentar la liberación de la hormona adrenocorticótropa, glucocorticoides y catecolaminas. La síntesis de testosterona se inhibe y aumenta su metabolismo hepático. Personas con alcoholismo crónico tienen signos de hipogonadismo y feminización (Martin PR., ND). Este conjunto de deficiencias biológicas da pie a que existan mutaciones y por consecuente enfermedades.

Este conjunto de relaciones muestra una relación directa a la susceptibilidad del paciente a padecer problemas médicos en sectores como metabólico, malnutrición, gastrointestinal, endocrino, neurológico y cardiovascular. Por lo que tomar los factores de la ingesta de tabaco y alcohol son determinantes para determinar si un paciente padecerá cáncer o no.

3.3 Implementación digital

Debido al crecimiento exponencial en avances tecnológicos de la última década junto con la creciente adopción de tecnología por parte de diferentes industrias se han podido desarrollar técnicas para incrementar la precisión medica al momento de hacer diagnósticos a los pacientes al igual que para poder automatizar este proceso a través de procesos como Deep Learning.

Han existido implementaciones de evaluación médica a través de técnicas como el uso de elementos radiológicos, citológicos para el cáncer de pulmón. En el estudio de la Mayo Clinic donde se involucraron 10933 pacientes ambulatorios con un alto índice de padecer cáncer de pulmón. En este estudio se logró únicamente detectar dos tercios de los canceres de pulmón con una radiografía de torax mientras que solo una quinta parte se logró mediante un examen citológico (S. Fontana et al., 2019). En este caso la radiografia resulto ser mucho mas eficaz para la detección de esta enfermedad mientras que un análisis sin esta resultó ser mucho mas ineficiente. De esta manera es posible demostrar como desde el uso más pequeño de tecnologías al ámbito medico es posible lograr un mejor análisis y junto con el conocimiento y expertise de un médico es posible llegar a un diagnostico acertado.

3.4 Aprendizaje automatizado e inteligencia artificial para uso médico

La inteligencia artificial y el Deep Learning se han usado para hacer modelos de análisis sumamente exactos en áreas como la economía, negocios, climatología, etc.

Dentro de la última década se ha comenzado a democratizar y popularizar el uso de estas tecnologías para el sector de la salud.

Casos como el de L1CAM el cual pudo realizar un excelente pronostico del cáncer endometrio tipo 1 en estadio 1. En este se realizó un estudio coherente multicéntrico retrospectiva para determinar la eficacia de L1CAM por inmunohistoquímica (S. Fontana et al., 2019). En este método se utilizaron métodos anidados de análisis tal como el Kaplan-Meier, el modelo de riesgo

proporcional de Cox para analizar la supervivencia y los análisis multivariables. Todos estos métodos fueron utilizados en conjunto en un modelo de aprendizaje automático para validar variables y poder predecir de esta manera la recurrencia y la muerte.

Han existido otros casos donde se ha hecho un trabajo colaborativo entre imágenes de radiografía, análisis por Deep Learning y análisis clínico tradicional. En el artículo An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening (2019) se realizó una recolección de datos de imágenes cervicales digitalizadas y archivadas de la detección, se alimento a un algoritmo a través de un proceso de capacitación y validación y posteriormente se imprimió 0-1 en caso de que se detectara cáncer. Este algoritmo detecto efectivamente los casos de cáncer/ pre cáncer acumulativos con mayor precisión que la interpretación original de un cervigramas o citología convencional. Mostrando asi que la utilización de procesos automatizados para la evaluación de imágenes cervicales masificadas en hospitales podría permitir la difusión de una detección de cáncer efectiva para que sea tratable y no existan casos de evolución de la enfermedad

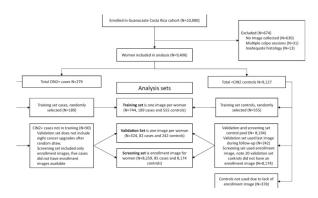


Diagrama del algoritmo de la inteligencia artificial (Hu L., Bell D. et al., 2019)

4. Inteligencia artificial

De acuerdo con Jannie Benyus la biomimética es un método por medio del cual los diseñadores e ingenieros hacen investigaciones biológicas para determinar cómo los organismos resuelven problemas complejos (Rocha Rangel, 2019).

Los científicos llevan siglos buscando comprender el funcionamiento del cerebro humano, a pesar de que aun no hay una noción completa de su funcionamiento y comportamiento se han logrado comprender partes de su funcionamiento y por consecuente se han logrado replicar. La inteligencia artificial funciona a través de modelos matemáticos, los cuales imitan la manera en la que el cerebro, específicamente las neuronas procesan la información.

Machine Learning funciona a través de Deep Learning (aprendizaje profundo) descubriendo modelos vastos y jerárquicos que representan distribuciones de probabilidad en conjuntos de datos (Goodfellow et al., 2018). De acuerdo con Christos Stergiou y Dimitrios Siganos (1996), una red neuronal es un paradigma de procesamiento de información basada en el sistema nervioso central humano. Está compuesto de elementos altamente interconectados llamados neuronas, las cuales trabajan en conjunto para resolver problemas. Las redes neuronales permiten extraer patrones y detectar tendencias con relaciones complejas para el ser humano.

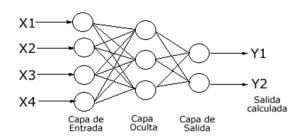


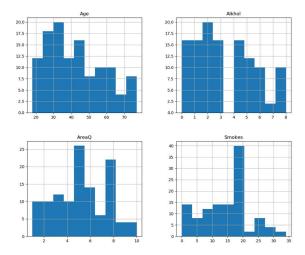
Diagrama de una red neuronal (De Antonio, 2019)

4.1 Planteamiento de la solución

Para poder tratar el cáncer es importante poder contar con un diagnostico adecuado, es decir que sea preciso, en una etapa temprana, rápido y sobre todo de bajo costo. En México como se mencionó anteriormente los casos de cáncer de pulmón son sumamente altos principalmente por los hábitos de la población, sin embargo, factores como la detección temprana también juegan un rol principal en estos casos. Por ello el utilizar un método que sea eficiente y que pueda obtener un diagnostico acertado rápidamente es de vital importancia para poder salvar la vida de millones de personas. Se desarrolló un modelo basado en clasificación binaria para poder determinar si un paciente es susceptible a padecer cáncer o no. Este modelo funciona como un método de prevención y de detección temprana de esta enfermedad evitando que se requiera de un análisis costoso, ineficiente y hasta cierto punto innecesario como los mencionados anteriormente.

4.2 Obtención de información médica

Para poder realizar el modelo fue necesario obtener una base de datos relevante para asi poder entrenar y probar la precisión, velocidad al igual que la eficacia del modelo. En este caso se obtuvo un registro medico con nombres editados y reemplazados por falsos para mantener la confidencialidad de los pacientes donadores de información. Este conjunto de datos se obtuvo de la página Kaggle del autor Yusuf Dede, ingeniero de software en Turkia. Este conjunto de datos contiene información de 59 pacientes.



Grafica 1.0 variables registradas de pacientes

Los datos están preprocesados a traves de un sistema de *Weights* o pesos. Es decir que a traves de una distribución binomial se establecen los máximos, mínimos y promedios para asi poder colocar la información del paciente y sustituirla por números mas de menor valor en una escala predefinida. Esto se realizo para las variables de edad, cantidad de tabaco consumido, área del pulmón, alcohol consumido y para el caso del resultado se estableció 1 o 0 en caso de que tenga o no cáncer el paciente.

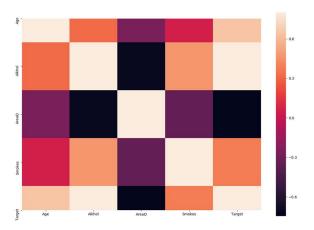
El conjunto de datos ha sido utilizado por 1748 modelos médicos con distintas

implementaciones para la predicción de cáncer de pulmón.

4.3 Procesamiento de información

Para el procesamiento de información se utilizaron métodos estadísticos para las neuronas. Las redes neuronales artificiales consisten en conjunto de unidades (nodos o neuronas), las cuales se interconecta para transmitirse señales tal como lo hace una neurona biológica, en este caso las predicciones y produciendo como valor de salida un 1 o un 0.

Como preprocesamiento y análisis de la información se realizo una grafica de tipo heatmap para poder ver la relación de las variables seleccionadas del conjunto de datos y analizar asi si existe una relación directa entre estas variables.



Grafica 1.1 heatmap para la correlación de variables

En la matriz ortogonal se aprecia que las coordenadas son proporcionales, coinciden mostrando una trasversa por lo que existe una relación proporcional directa. A partir de este análisis se pudo determinar que el conjunto de datos era el adecuado para poder predecir si un paciente padece de cáncer o

no. Esto con su debido entrenamiento ya que en el caso de esta investigación se utilizó en un conjunto de datos especifico no fue exportado a nuevos datos.

5. Creación del modelo

El modelo fue creado en Python 3.7.6, para la carga de la base de datos y el preprocesamiento se utilizó la librería de Pandas, esto ya que permite captar grandes cantidades de datos y manipularlas libremente. Para la base del modelo se utilizó Keras y Tensorflow, utilizando el modelo secuencial y el de densidad, esto para poder medir los datos y poder alimentarlos en el modelo clasificador. Posteriormente para validar la información procesada por las neuronas, establecer etiquetas y otros elementos para agilizar el proceso de análisis básico de los datos se utilizó SkLearn.

El algoritmo funcionó de la siguiente manera; se cargan los datos del archivo csv a un dataframe, se dividen los datos en variables X y Y. Para este caso los pacientes fueron las Y y todos los atributos fungieron como X. Posteriormente se hizo uso de un encoder para las etiquetas de cada columna y se transformaron los datos a un conjunto. Posteriormente se creo el modelo utilizando un modelo secuencial de red neuronal tomando 3 capas, la primera con un activador con función sigmoid, la segunda con una capa Relu y la tercera y cuarta con un activador Relu. Se compilaba el modelo en un clasificador binario con Crossentropy con un optimizador Adam. Estos datos luego eran metidos en un crossvalidator para evaluar el modelo de aprendizaje automático en este dataset.

5.1 Tiempo de ejecución y comportamiento del modelo

El modelo tarda un aproximado de entre 1 a 3 minutos para un batch size de 5 con epochs de 100. Es decir que toma 5 datos de la base de datos, las compara, analiza y finalmente categoriza cien veces con el conjunto de datos para alimentar la neurona y posteriormente obtener un resultado binario. Esto es un comportamiento similar a como una neurona convencional analiza la información, esta recibe la información, la analiza y posteriormente la reenvía a otras neuronas.

El modelo tuvo una precisión de entre 87.14% (20.40%) y 97.33% (15.28%) lo que indica que este modelo categoriza adecuadamente los datos y en caso de que se vea presentado a datos no conocidos podrá ser capaz de categorizar al paciente, dándole asi un diagnostico preventivo respecto a sus hábitos, edad y tamaño del pulmón. La modelo varia en porcentaje de precisión debido a la pequeña cantidad de datos por lo que si este modelo fuese entrenado en una base de datos con una cantidad significativa de información de pacientes seria capaz de estabilizar sus niveles de precisión.

El código del proyecto esta disponible para su utilización libre y Open Source en el siguiente link:

https://github.com/SeaWar741/ITC/tree/mas ter/1er Semestre/Sistemas%20Biologicos

6. Conclusiones

A través de este modelo de inteligencia artificial fue posible demostrar el uso de un clasificador binario para la detección temprana de cáncer pulmonar. Mostrando que es posible utilizar estas tecnologías en el ámbito médico como una herramienta clínica de bajo costo, eficiente y sobre todo preventiva. Esta investigación también demostró que en algunos casos el error humano es catastrófico no solo en un caso sino en la mayoría por lo que combinar la tecnología biomimetica a través de la inteligencia artificial con los conocimientos de un medico pueden ser el mejor caso para detectar, tratar y sobre todo superar esta enfermedad. La implementación de estos modelos en la industria medica han comenzado a utilizarse por lo que democratizar estas tecnologías es de vital importancia para continuar con el desarrollo de curas y tratamientos para el cáncer pulmonar.

8, Trabajo futuro

Este trabajo ha permitido explorar distintas disyuntivas del cáncer por lo que a futuro este modelo debería de ser probado en conjuntos de datos más grandes, deberían de realizarse comparaciones entre modelos de predicción y se podrían desarrollar modelos de este estilo para desarrollo de curas para el cáncer a través de redes neuronales, las cuales como se han demostrado en el trabajo son una herramienta sumamente eficaz y precisa.

Referencias

Cáncer. (2019). Retrieved 25 October 2019, from https://www.who.int/topics/cancer/es/

Chesnutt, Asha N., et al..(2019)"Screening for Lung Cancer." Current Medical Diagnosis and Treatment 2020 Eds. Maxine A. Papadakis, et al. New York, NY: McGraw-Hill, , http://0-

accessmedicine.mhmedical.com.millenium.i tesm.mx/content.aspx?bookid=2683§io nid=225058361.

De Antonio, J. (2019). Redes de Neuronas Artificiales con NetBeans. Retrieved 26 October 2019, from https://unpocodejava.com/2011/03/29/redesde-neuronas-artificiales-con-netbeans/

Dede, Y. (2019). Lung Cancer DataSet. Retrieved 26 October 2019, from https://www.kaggle.com/yusufdede/lungcancer-dataset

Dela Cruz, Charles S., et al..(2015)."Epidemiology of Lung Cancer." Fishman's Pulmonary Diseases and Disorders, Fifth Edition Eds. Michael A. Grippi, et al. New York, NY: McGraw-Hill, 2015, http://o-accessmedicine.mhmedical.com.millenium.itesm.mx/content.aspx?bookid=1344§ionid=81197135.

Dube S, McClave A, James C, Caraballo R, Kaufmann R, Pechacek T. (2009). Vital Signs: Current cigarette smoking among adults aged > = 18 years – United States, 2009. In: Office on Smoking and Health. CDC., ed. MMWR Morb Mortal Wkly Rep. 2010;59:1135–1140. [PubMed: 20829747] Martin PR. (ND). Substance-Related and Addictive Disorders. In: Ebert MH, Leckman JF, Petrakis IL. eds. Current Diagnosis & Treatment: Psychiatry, 3e New York, NY: McGraw-Hill; . http://0accessmedicine.mhmedical.com.millenium.i tesm.mx/content.aspx?bookid=2509§io nid=200808792. Accessed October 25, 2019.

EFE. (2019). El 99 % de casos de cáncer de pulmón en México se diagnostican tardíamente. Retrieved 26 October 2019, from

https://www.efe.com/efe/america/mexico/el-99-de-casos-cancer-pulmon-en-mexico-se-diagnostican-tardiamente/50000545-3945535

Goodfellow,I., Pouget-Abadie*, J.,Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S. et al. (2018). *Generative Adversarial Nets*. Recuperado

de_https://arxiv.org/pdf/1406.2661.pdf

Hu L., Bell D., Antani S., Xue Z., Yu K., Horning M., Gachuhi N., Wilson B., Jaiswal, Brian Befano M., Long L., Herrero R., Einstein M., Burk R., Demarco M., Gage J., Rodriguez A., Wentzensen N., Schiffman M..(2019). An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening, *JNCI: Journal of the National Cancer Institute*, Volume 111, Issue 9, September 2019, Pages 923–932, https://doi.org/10.1093/jnci/djy225

INEGI. (2019). ESTADÍSTICAS A PROPÓSITO DEL... DÍA MUNDIAL CONTRA EL CÁNCER (4 DE FEBRERO). Retrieved 26 October 2019, from https://www.inegi.org.mx/contenidos/salade prensa/aproposito/2018/cancer2018_Nal.pdf

Mattson ME, Pollack ES, Cullen JW. (1987). What are the odds that smoking will kill you? Am J Public Health. 1987;77(4):425–431. [PubMed: 3826460]

NIH. (2019). Diccionario de cáncer. Retrieved 26 October 2019, from https://www.cancer.gov/espanol/publicaciones/diccionario/def/metastasis

Pisani P, Bray F, Parkin DM. (2002) Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. Int J Cancer. 2002;97(1):72–81. [PubMed: 11774246]

Research, U. (2019). How cancer starts. Retrieved 26 October 2019, from https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts

Rocha Rangel, E. (2019). Biomimética: de la naturaleza a la creación humana - Revista Ciencias. Retrieved 26 October 2019, from https://www.revistaciencias.unam.mx/es/99-revistas/revista-ciencias-98/611-biomimetica-de-la-naturaleza-a-la-creacion-humana.html

S. Fontana, R., R. Sanderson, D., F. Taylor, W., B. Woolner, L., Miller, W., & R. Muhm, J. et al. (2019). Early Lung Cancer Detection: Results of the Initial (Prevalence) Radiologic and Cytologic Screening in the Mayo Clinic Study | American Review of Respiratory Disease. Retrieved 26 October 2019, from

https://www.atsjournals.org/doi/abs/10.1164/arrd.1984.130.4.561

Saínz Menéndez, Benito. (2006). Tumores benignos y malignos del pulmón: Clasificación. Diagnóstico. Tratamiento. *Revista Cubana de Cirugía*, 45(3-4) Recuperado en 25 de octubre de 2019, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0034-74932006000300021&lng=es&tlng=es.

Smith CJ, Perfetti TA, Rumple MA, Rodgman A, Doolittle DJ.(2000) "IARC group 2A Carcinogens" reported in cigarette mainstream smoke. Food Chem Toxicol. 2000;38(4):371–383. [PubMed: 10722891]

Stephen S. Hecht, Tobacco Smoke Carcinogens and Lung Cancer, JNCI: Journal of the National Cancer Institute, Volume 91, Issue 14, 21 July 1999, Pages 1194–1210, https://doi.org/10.1093/jnci/91.14.1194

Tumores benignos: MedlinePlus en español. (2019). Retrieved 26 October 2019, from

https://medlineplus.gov/spanish/benigntumo

rs.html