FORMATO DE APLICACIÓN DE FERIA DE PROYECTOS FIEECS 2018-I

DESARROLLO DE UN MODELO BASADO EN MACHINE LEARNING PARA LA PREDICCION E IDENTIFICACION DE FACTORES IMPORTANTES DEL CANCER DE MAMA

Integrantes: Tinco Dominguez, Alfredo Walter

Docente: Amelida S Pinedo/ Carlos R. Franco

Asignatura: Taller de elaboración de tesis

Escuela Profesional de Ingeniería Estadística

Facultad de Ingeniería Económica, Estadística y CCSS

Universidad Nacional de Ingeniería

RESUMEN

En este trabajo presentamos un análisis predictivo utilizando el machine Learning para poder predecir si un tumor analizado tiene calificación de benigno o maligno.

Utilizaremos 3 métodos de aprendizaje automático: KNN (k vecinos más cercanos) Xgboost(gradiente) y el SVM(máquina de soporte vectorial).

El poder de análisis y predicción de estos métodos automatizados son de gran potencia, lo importante de estos son que aprenden de manera propia, analizando la data con una muestra de entrenamiento y la evalúa en una data de testeo.

Asu vez, utilizaremos una base de datos brindada, con registros de características fisiológicas de tumores extraídos de los senos de mujeres, con esto lograremos identificar utilizando los métodos predictivos, cuál de estos predice mejor y que factores son importantes para determinar si un tumor es benigno o maligno.

INTRODUCCIÓN

El cáncer de mama es una de las patologías malignas más comunes en el Perú y ataca principalmente a mujeres. El Instituto Nacional de Enfermedades Neoplásicas es un centro de referencia para el diagnóstico y tratamiento del cáncer en general. Siendo está muy común, es importante determinar eficientemente los casos de "tumores malignos" (cáncer) para su pronto tratamiento de acuerdo al diagnóstico, por lo que se requiere un modelo que permita predecir el diagnóstico de cáncer de mama.

"El cáncer de mama ha cambiado de perfil. En los últimos años, en el Perú, la incidencia ha aumentado y afecta a 42 de cada 100,000 habitantes, pero ya no solo se presenta entre las mujeres a partir de los 40 años, sino que se detecta a edades más tempranas, desde los 30 años o incluso desde los 25 años, advirtió el cirujano oncólogo de la Liga contra el Cáncer Marco Velarde.

El experto aseveró que en esta enfermedad el Perú está a la par de los países occidentales más avanzados, que son justamente los que tienen más incidencia y más mortalidad por cáncer de mama. Según la Liga Contra el Cáncer, en el país, cada año se presentan unos 5,000 casos nuevos y la mayoría llega en estadios muy avanzado, cuando las probabilidades de curación son de solo el 50%. Se estima que cada año unas 2,000 mujeres por esta enfermedad.

Velarde señaló que un cáncer de mama detectado a tiempo tiene el 95% de probabilidades de curación. Por ello, recomendó a las mujeres hacerse un examen clínico anual, una mamografía a partir de los 40 años y una ecografía de mama desde los 25 años, para detectar precozmente alguna neoplasia.

"Cada vez tenemos población más joven con cáncer. Aunque la mayoría de casos son mujeres mayores de 40 años, los porcentajes de pacientes que a los 30 años o a los 25 años tienen cáncer de mama están aumentando. Por eso la valla de prevención también está bajando", dijo.

El médico refirió que aún no hay estudios definitivos de por qué se estarían presentando más casos de cáncer de mama entre población joven. "Un factor tiene que ver con lo genético, pues cada vez hay más población que tiene familia con cáncer y genéticamente las personas están predispuestas a desarrollar la neoplasia. Si una mujer tiene cáncer a los 45 años probablemente sus hijas tendrán la enfermedad a los 40 años y las hijas de sus hijas a los 35 años", manifestó.

PRESENTACIÓN DEL PROBLEMA

En el Perú, el riesgo de morir para las personas que tienen cáncer es sumamente alto, más que el promedio de Sudamérica y casi el doble que en Estados Unidos, según un estudio de The Economist. La investigación, revelada en el Roche Press Day de Buenos Aires, evaluó las fortalezas y debilidades en el control de esta enfermedad en la región.

Según la Unidad de Inteligencia del semanario británico, Perú tiene una tasa mortalidad sobre incidencia de 0.60, mientras que el promedio de Sudamérica es de 0.53 y el de Estados Unidos 0.33 (Europa tiene 0.40). Mientras menor sea el número, mayor es la eficacia de las políticas públicas y de los programas de prevención del cáncer. Otros ejemplos son Chile y Argentina, con 0.59 y 0.55 respectivamente. El más bajo de la región es Bolivia, con 0.67.

Este indicador muestra cuántos pacientes mueren en relación a cuántos casos nuevos de cáncer hay cada año.

Cabe recordar que el cáncer es la primera causa de muerte en el Perú: representa el 19% del total de decesos. La incidencia es de 154,5 casos cada 100 mil habitantes.

En Latinoamérica, el cáncer es la segunda principal razón de fallecimientos. Entre el 60% y 70% de los pacientes son diagnosticados en estadios avanzados. Se espera que para el 2035 las muertes por cáncer se dupliquen.

El informe "Control del cáncer, acceso y desigualdad en América Latina: una historia de luces y sombras", que analizó las realidades de Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, México, Panamá, Paraguay, Perú y Uruguay, señala que este país presenta deficiencias en la financiación y en la disponibilidad de medicamentos, radioterapia y cuidados paliativos, capacitación de personal de salud en atención primaria, descentralización de la atención médica, entre otros aspectos más generales como el simple acceso a la salud.

Es estrictamente saber del control que uno debe hacerse o instar a que se hagan los respectivos controles para la prevención que se tiene que hacer frente a este problema pues como vemos en descripciones anteriores, el cáncer de mama en el Perú esta cobrando un gran impacto, ahora ya no solo se detectan casos en mujeres sino también en varones, en menores casos, pero se presentan. La detección de este problema a tiempo reduce el riesgo de una muerte por cáncer de mama. Como un inicio a esta prevención, analizar las características fisiológicas de los tumores extraídos y así obtener resultados de diagnóstico utilizando modelos de Machine Learning, servirían como base para diferentes formas de detección de cáncer de mama.

OBJETIVOS

Encontrar el modelo de machine Learning que prediga eficientemente el diagnóstico de cáncer de mama.

Determinar si el modelo de XGboost de machine Learning predice mejor el diagnóstico de cáncer de mama.

Determinar los factores importantes para el mejor modelo predictivo.

DESCRIPCIÓN DE LA SOLUCIÓN

Como podemos notar en el gráfico adjunto, el procedimiento de tratamiento de la data para poder analizarlo es de prioridad. Tenemos en primera instancia que el target estaba calificado como Malgino y Benigno, que para nuestra conveniencia y por los métodos a utilizar tenemos que cambiarla, codificándola con valores 0 y 1. Luego de esto necesitamos limpiar los datos miseados o atípicos, que en este análisis no llegaron a encontrarse.

Luego podremos realizar la visualización del comportamiento de las variables, podremos realizar gráficos de dispersión, gráficos de barras y en esta ocasión presentamos el diagrama de violín, que lo que hace es graficar a las variables con su respectiva distribución y un diagrama de cajas incrustado. Esto para ver sii alguna de las variables es significativa o influyente en el target.

Luego hacemos la particion de la data, un 0.7para el training y 0.3 para test data. Luego entrenamos el modelo y con el test data podremos ver cuán eficiente, o cuan predictivo es el modelo.

Finalmente para la comparación de los modelos utilizados, usaremos el indicador de accuracy, que nos indica cuantos porcentaje ha sido predicho.

RESULTADOS

De los modelos llegamos a obtener lo siguiente:

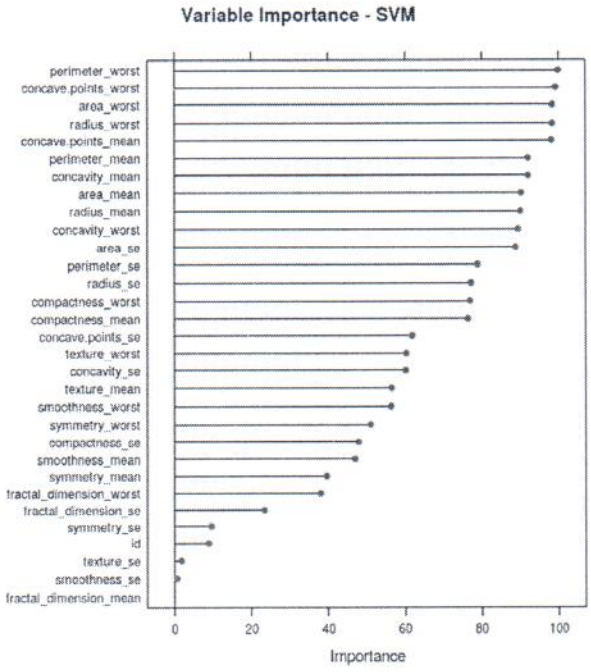
	score	accuracy
xgboost	0.937835	0.970588
KNN	0.393261	0.729412
SVM	0.949562	0.976471

CONCLUSIONES

Podemos concluir que las herramientas de machine Learning si nos ayudan a determinar modelos predictivos, así como factores importantes para cada uno de estos. En esta oportunidad llegamos a inferir mediante la capacidad de predicción del modelo que el SVM predice mucho mejor que todos los modelos, también nos proporciona un score muy alto.

En tanto a los factores importantes en el modelo del SVM, tenemos que el perímetro, puntos cóncavos, el área mala, el radio malo son factores importantes en este modelo determinantes para la identificación de cáncer de mama.

Hablando de los factores importantes:



El perímetro, el área y los puntos cóncavos son importantes.

RECOMENDACIONES

Se recomienda hacer este mismo estudio con poblaciones de edades diferenciadas, así también usando la técnica de ensamblado para poder determinar una mejor predicción de los casos de tumores que sean benignos y malignos.

BIBLIOGRAFÍA

1. Sung Gwe A, Hak Min L, Sang Hoon Ch, Seung Ah L, Seung Hyun H, Joon J, *et al.* Prognostic factors for patients with bone-only metastasis in Breast Cancer. Yonsei Med J. 2013;54(5):116877.

2 . Phillips KA, Milne RL, Friedlander ML, Jenkins MA, McCredie MR, Giles GG, *et al.* Prognosis of premenopausal breast cancer and childbirth prior to diagnosis. J Clin Oncol. 2004;22(4):699-705.

3 . Kroman N, Wohlfahrt J, Andersen KW, Mouridsen HT, Westergaard T, Melbye M. Time since childbirth and prognosis in primary breast cancer: population based study. BMJ. 1997;315(7112):8515.

4. Early Breast CancerTrialists' Collaborative Group, McGale P, Taylor C, Correa C, Cutter D, Duane F, *et al.* Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8135 women in 22 randomised trials. Lancet. 2014;383(9935):2127-35.

5. Hoffman HJ, Khan A, Ajmera KM, Zolfaghari L, Schenfeld JR, Levine PH. Initial response to chemotherapy, not delay in diagnosis, predicts overall survival in inflammatory breast cancer cases. Am J Clin Oncol. 2014;37(4):315-21.

6. Early Breast CancerTrialists' Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. Lancet. 2000;355(9217):1757-70.

7 . Yaming Li, Meena M, Qiang H, Qifeng Y, Bruce H. Post-mastectomy radiotherapy for breast cancer patients with T1-T2 and 1-3 positive lymph nodes: a meta-analysis. PLoS One. 2013;8(12):81765.

8. Mastering Machine Learning with python

9. https://archive.ics.uci.edu/ml/datasets.html