Juan Miguel Gutierrez Vidal

Alejandra Campo Archbold

Juan Fernando Pérez Bernal

Summer School Big Data 2018

**CLASIFICACION DE RIESGO CANCER CERVICAL**

Se aplicaron diferentes métodos algoritmos de selección para predecir el nivel de riesgo que una mujer posea cáncer Cervical dado que se le detecto en 4 test que existen para detectar cáncer cervical: Hinselmann, Schiller, Biopsia, Citología . Así mismo se analizo que variables son las más importantes para la predicción del nivel de riesgo. Por último, se analizó la diferencia de precisión de cada modelo utilizando las curvas de *Precision-Recall*

El problema reside en reconocer el mejor modelo que sea capaz de ajustarse mejor a los datos para predecir el nivel de riesgo de sufrir de cáncer cervical dado las variables recolectadas. Entre los diferentes modelos que se utilizaron fueron LR (Regresión Lineal), PCA (Principal Componentes), QDA (Discriminante Cuadrático), LDA (Discriminante Linear), VSC (Clasificación Vectorial), Arboles de Decisión.

La base de datos poseía 858 datos de mujeres que se hicieron diferentes test de cáncer cervical y respondieron una encuesta en Caracas, Venezuela. El cáncer de cuello uterino es el tercer tipo de cáncer más común en mujeres. Las diferentes variables que se utilizaron para predecir el modelo fueron la edad, numero de compañeros sexuales, edad al primer encuentro sexual, si fuma y cuantas cajetillas fuma al año; si toma anticonceptivos hormonales y por cuanto los lleva utilizando; si utiliza IUD y por cuanto tiempo lo lleva utilizando. Por otra parte, también se les analizo si poseían enfermedades relacionadas con relaciones sexuales como condilomatosis, sífilis, herpes, VIH.

Primero, para analizar el riesgo de poseer cáncer cervical se creó una nueva variable llamada *“Cancer Risk”* la cual fue definida:

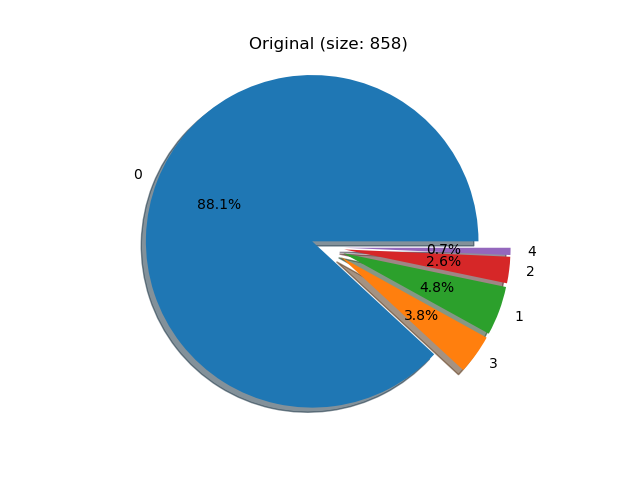
Donde toma un valor de 0 hasta 4, donde 0 donde su riesgo es mínimo o nulo, hasta ascender a 4 donde el riesgo es muy alto y ya posiblemente el cáncer este expandido por todo el cáncer. El porcentaje de personas por cada nivel de riesgo se puede ver en la Figura 1.

Figura 1. De las 858 mujeres los riesgos se distribuían:

Nivel 1: 44 Personas

Nivel 2: 22 Personas

Nivel 3: 33 Personas

Nivel 4:6 Personas

Después fue necesario balancear la base de datos para que el aprendizaje de maquina no presentara inconsistencias.

Una vez hecha la clasificación se utilizó el método Precisión- Exhaustividad(“precisión-recall”) donde la precisión trata de responder que proporción de identificaciones positivas fue correcta. Mientras la exhaustividad mide que tantos positivos reales se identifico correctamente. Para evaluar los modelos de clasificación se debe mirar la precisión y recuperación(exhaustividad) del modelo. Si a medida que aumenta la exhaustividad la precisión aumenta, significa que el modelo está devolviendo resultados más exactos/precisos.

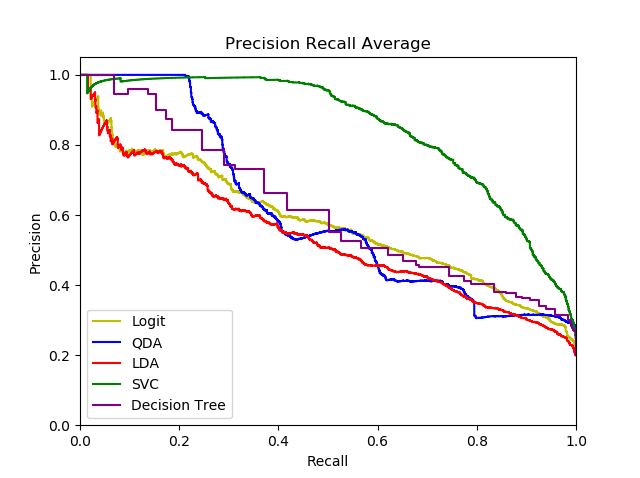
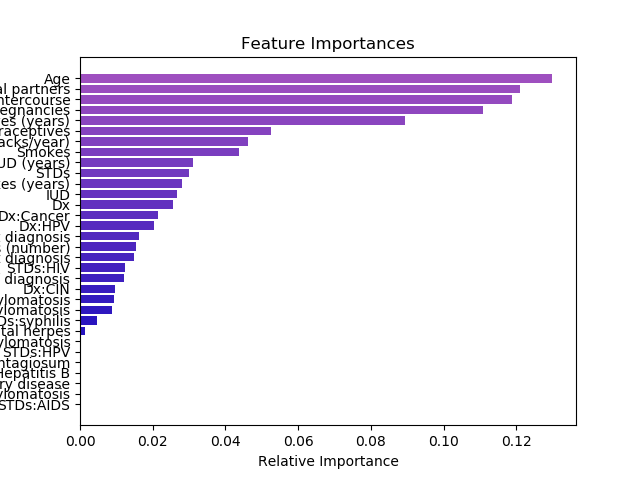


Figura 2. Grafica de Precision-Recall de los modelos para el promedio de todas las clases. “SVC=SVM”

juntas

En la Figura 2 se puede ver que el mejor modelo de clasificación es SVM (Maquina de vectores de soporte) .Es mejor prediciendo que los demás modelos, este seguramente se debe a que es capaz de clasificar mejor los niveles de riesgo 2,3,4 que poseen muy pocas observaciones. En las matrices de confusión adjuntadas, se puede llegar a la misma conclusión, donde SVM tiene una precisión de 75%. En promedio mas alto un 20% que los demás.

Por otra parte, se analizó cuáles eran las variables que más poder de predicción poseían en los modelos.



La edad es la primera de las variables que mas poder predictivo posee, esto es posible dado que a medida que pasan más anos es más probable que se hayan presentado situaciones donde se transmitió cáncer. Esto puede relacionarse que a medida que pasan los años hay mayores compañeros sexuales, y mediante estos se haya absorbido la semilla del cáncer. Por ultimo la tercera variable mas importante es cuando la mujer presencia su primer encuentro sexual, esto puede deberse a que a medida que mas promiscuan sean, mayor son las relaciones sexuales que mantienen y por lo tanto mayores probabilidades de que hayan obtenido el cáncer. Estas tres primeras variables indican que las relaciones sexuales y el cuidado de las mujeres esta muy correlación con sufrir cáncer cervical. Por otro lado, es sorprendente darse cuenta de que poseer otras enfermedades de contagio sexual no tiene mucha relación con sufrir de cáncer cervical, donde se tendería a pensar que sí.

En conclusión, a pesar de que la base de datos tuviera pocas muestras, se puede evidenciar que el modelo de clasificación vectorial es por mucho el mejor modelo para predecir comparado con los demás. No solo esto, si no que fue posible llegar a analizar que las variables que mas afectan que una mujer pueda contraer el cáncer, están relacionadas en general con la edad, primer encuentro sexual y compañeros sexuales.

# Bibliografía

Google. (s.f.). Obtenido de https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Learning, S. S. (s.f.). Obtenido de https://www.kaggle.com/loveall/cervical-cancer-risk-classification

SaurabhHarsh. (s.f.). *Kaggle*. Obtenido de https://www.kaggle.com/saurabhharsh/financial-distress-prediction

Skicit-Learn. (s.f.). *Skicit-Learn*. Obtenido de http://scikit-learn.org/stable/auto\_examples/model\_selection/plot\_precision\_recall.html

ANEXOS.

