

Saturdays.AI Guadalajara

Equipo 3c (Cobra Kai)

Participantes:

Lisset Valenzuela

Rodrigo Caballero

Román Rodríguez

Jarumi Cruz Leos

Diana López

Magdalena Loaeza

Ricardo Esparza

Puntos generales del proyecto en Saturdays.ai

***1. Nombre del proyecto***

**Algoritmo de Machine Learning para predicción de un ataque cardíaco**

***2. Eje rector del proyecto***

Sector salud - Médicos

Debido a que no hay mecanismos automáticos que permitan prever afecciones cardiacas considerando factores clínicos de manera eficiente, no invasiva y óptima permite una predicción del diagnóstico del paciente con una asertividad mayor del 90% usando algoritmos de Machine Learning.

***3. Planteamiento del problema***

Detectamos un alto índice de afecciones cardiacas que pudieran haberse previsto si se hubiera puesto atención a la sintomatología del paciente y a sus resultados clínicos, por ejemplo:

* Presión, opresión, dolor, o sensación de compresión o dolor en el pecho o en los brazos, que puede propagarse hacia el cuello, la mandíbula o la espalda.
* Náuseas, indigestión, ardor de estómago o dolor abdominal
* Falta de aire
* Sudor frío
* Fatiga
* Aturdimiento o mareos repentinos

Con esta información y el análisis clínico un médico general puede realizar diagnósticos parecidos a los de un especialista.

***4. Descripción de la solución a la problemática detectada.***

Con base a los registros históricos de casos existentes y síntomas se pretende prever o anticipar el riesgo de sufrir una afección cardíaca.

Debido a que no hay suficientes cardiólogos o especialistas que atiendan a esta gran demanda de casos, con esta solución es posible que un médico general pueda utilizar este algoritmo y realizar una búsqueda de los resultados clínicos y utilizar el algoritmo para predecir en ese momento la probabilidad de un ataque cardíaco.

En este caso se plantea una solución que derivada de una consulta general, sin tener un estudio clínico como el electrocardiograma se lograr una mayor precisión en la predicción, los datos que pueden considerarse son:

* **age:** Edad de la persona
* **sex:** Género (1 = masculino, 0 = femenino)
* **cp:** Dolor torácico experimentado (1= angina típica, 2=angina atípica, 3=dolor no anginal, 4=asintomático)
* **trestbps:** Presión arterial en reposo de la persona (mm Hg al ingreso al hospital)
* **chol:** Nivel de colesterol en mg/dl

***5. Hipótesis ¿Cuál es tu hipótesis a probar?***

¿Puede un algoritmo de machine learning lograr una predicción de diagnóstico de un ataque cardiaco con base a los síntomas del paciente con una mayor precisión?

***6. Metodología***

**a. Descripción de los datos**

* **age:** Edad de la persona
* **sex:** Género (1 = masculino, 0 = femenino)
* **cp:** Dolor torácico experimentado (1= angina típica, 2=angina atípica, 3=dolor no anginal, 4=asintomático)
* **trestbps:** Presión arterial en reposo de la persona (mm Hg al ingreso al hospital)
* **chol:** Nivel de colesterol en mg/dl
* **fbs:** Nivel de azúcar en sangre en ayunas (> 120 mg / dl, 1 = verdadero; 0 = falso)
* **restecg:** Medición electrocardiográfica en reposo (0 = normal, 1 = con anormalidad de la onda ST-T, 2 = mostrando hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)
* **thalach:** Frecuencia cardíaca máxima alcanzada.
* **exang:** Angina inducida por el ejercicio (1=si; 0=no)
* **oldpeak:** Depresión de ST inducida por el ejercicio en relación con el descanso ('ST' se relaciona con las posiciones en la gráfica del ECG)
* **slope:** Pendiente del segmento ST del ejercicio pico (1=pendiente ascendente, 2=plano, 3=pendiente descendente)
* **ca:** Número de vasos principales (0 - 3)
* **thal:** Trastorno sanguíneo llamado talasemia (3 = normal; 6 = defecto fijo; 7 = defecto reversible)
* **target:** Enfermedad cardiac (0=No, 1=Sí)

**b. Descripción del modelo a utilizar**

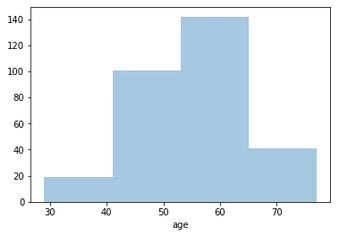
Se utilizará un Gradient Boosting Classifier al tratarse de un modelo para predicción de clasificación binaria (1,0) de la librería sklearn de Python.

**c. Delimitaciones**

**i. De datos**

No se obtuvo el volumen deseado para obtener muestras significativas

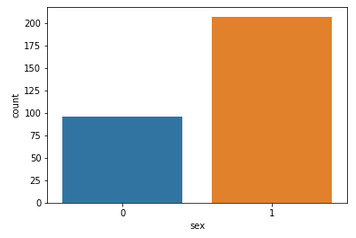
La delimitación de los datos tiene un rango entre 41 y 65



Más de 2/3 de la base son hombres

Porcentage of males: 68.317

procentage of females: 31.683



**ii. Económicas**

Actualmente es muy costoso en términos de tiempo, capacidad de computo y necesidad de un especialista para poder realizar el diagnóstico; con esta propuesta se reduce el tiempo (diagnóstico), y un médico general puede realizar esta actividad.

Dentro de los datos seleccionados existe el análisis de colesterol que requiere realizarse previamente.

**iii. Temporales**

N/A

**iv. Teóricas**

Por el momento sólo se llega a un nivel de predicción hasta del 87 en la precisión.

***7. Marco Teórico Estudios elaborados anteriormente sobre la temática***

Machine learning para predicción cardiovascular.

Subestudio a partir de los datos del estudio MESA (*Multi-Ethnic Study of Atherosclerosis*), que prueba el *machine learning* (aprendizaje automático), como estrategia de *big data* puede ser útil para caracterizar el riesgo cardiovascular, predecir los resultados e identificar biomarcadores en los estudios poblacionales.

El objetivo que se planteó fue probar la capacidad de los *random survival forests* (RF), en castellano ‘bosques aleatorios de supervivencia’, una técnica de *machine learning*, para predecir seis resultados cardiovasculares en comparación con las escalas de riesgo cardiovascular estándar.

Los métodos empleados fueron los siguientes: se incluyeron participantes del estudio MESA. Las medidas basales se utilizaron para predecir los resultados cardiovasculares a lo largo de 12 años de seguimiento. MESA fue diseñado para estudiar la progresión de la enfermedad subclínica a eventos cardiovasculares, donde los participantes estaban inicialmente libres de enfermedad cardiovascular. Se incluyeron 6.814 participantes de MESA, de 45 a 84 años, de 4 etnias y 6 centros de todo Estados Unidos. Se obtuvieron 735 variables de técnicas de imagen y pruebas no invasivas, cuestionarios y paneles de biomarcadores. Utilizamos la técnica RF para identificar los 20 predictores principales de cada resultado.

En cuanto a los resultados las técnicas de imagen, la electrocardiografía y los biomarcadores en suero se encontraban altamente representadas en la parte superior de las listas de 20 más importantes predictores, en comparación con los factores de riesgo cardiovascular tradicionales. La edad fue el predictor más importante para la mortalidad por todas las causas. Los niveles de glucosa en ayunas y las medidas de ecografía carotídea fueron predictores importantes del ictus. El *score* de calcio coronario fue el predictor más importante de la enfermedad coronaria y todos los resultados combinados de la enfermedad cardiovascular aterosclerótica. La estructura y la función ventricular izquierda y la troponina-T cardiaca se encuentran entre los principales predictores de la incidencia de insuficiencia cardiaca. La creatinina, la edad y el índice tobillo-brazo se encontraban entre los principales predictores de fibrilación auricular. El factor de necrosis tisular-α y los receptores solubles de interleuquina-2, y los niveles de NT-proBNP fueron importantes en todos los resultados. La técnica de RF se comportó mejor que las escalas de riesgo establecidas mostrando una mayor precisión de predicción (disminuyendo el Brier *score* en un 10-25%).

Conclusiones: el *machine learning* junto con el *deep phenotyping* (‘fenotipado profundo’) mejora la precisión predictiva de eventos cardiovasculares en una población inicialmente asintomática. Estos métodos pueden conducir a una mayor comprensión de los marcadores de enfermedad subclínica sin asunciones apriorísticas de causalidad.

## Comentario

La predicción de eventos ha sido la piedra angular de la epidemiología cardiovascular, como ejemplifica el estudio de Framingham y otros estudios prospectivos que funcionan como pilares para gran parte de lo que incluye la medicina cardiovascular actual. Un objetivo fundamental de tales esfuerzos ha sido la predicción de eventos durante periodos de tiempo relativamente largos tales como como diez años o toda la vida restante. Estos esfuerzos nos han permitido caracterizar procesos de enfermedades subclínicas y objetivos de factores de riesgo clave para su modificación (por ejemplo, dejar de fumar, terapia con estatinas, control de la presión arterial). Los estudios epidemiológicos utilizados para derivar tales modelos predictivos con frecuencia contienen cientos o miles de variables. Es en este contexto que los métodos de *machine learning* pueden ser útiles como un medio para identificar los mejores predictores de resultados entre millones de puntos de datos fenotípicos.

Las técnicas de *machine learning*, como las técnicas aleatorias de supervivencia, pueden ser una metodología estadística efectiva para manejar datos biomédicos de mayor volumen, velocidad y variedad. Estos métodos no requieren suposiciones *a priori* con respecto a la causalidad y pueden ser adecuados definiendo el papel de los nuevos biomarcadores en la predicción de las enfermedades cardiovasculares.

Este artículo añade nueva información de gran interés, destacando:

* Los métodos de *machine learning* pueden ser más adecuados para una predicción de riesgo significativa en estudios epidemiológicos fenotípicamente extensos a gran escala respecto a los modelos de riesgos proporcionales de Cox o las escalas de riesgo.
* Los RF pueden ser una estrategia efectiva de machine learning para la predicción incidental de eventos cardiovasculares y la estratificación de riesgo en poblaciones grandes con grandes conjuntos de datos fenotípicos.

***8. Resultados***

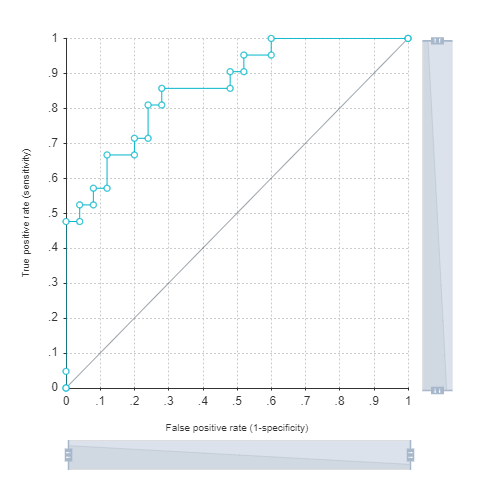
De acuerdo a las pruebas realizadas utilizando todos los factores antes mencionados y acotando aquellos que no requieren de un estudio clínico se pudo comprobar el aumento en la precisión y en la reducción de costos.

1er prueba 73%

2da prueba 78%

3er prueba 87%

Nuestro objetivo dado el modelo de predicción, es maximizar la ROC AUC y obtener las probabilidades de pertenecer a los grupos de clasificación



Buscamos una mayor Sensitivity por sobre la Specificity, debido a poder tomar acciones predictivas dado un resultado positivo y descargar en estudios superiores un falso negativo

|  |  |  |  |
| --- | --- | --- | --- |
| **Observed** | **Predicted** | | |
| **0** | **1** | **Percent Correct** |
| 0 | 15 | 6 | 71.4% |
| 1 | 5 | 20 | 80.0% |
| Percent Correct | 75.0% | 76.9% | 76.1% |

La área bajo la curva presenta buenos resultados, y más con una estrategia de Cross-Validation

|  |  |  |
| --- | --- | --- |
|  | **Holdout Score** | **Cross Validation Score** |
| Accuracy | 0.761 | 0.844 |
| Area Under ROC Curve | 0.853 | 0.928 |
| Precision | 0.769 | 0.840 |
| Recall | 0.800 | 0.885 |
| F1 Measure | 0.784 | 0.861 |
| Average Precision | 0.877 | 0.932 |
| Log Loss | 0.478 | 0.351 |

***9. Conclusiones finales y siguientes pasos***

La conclusión es que el algoritmo soluciona eficientemente una predicción con un menor esfuerzo médico, económico y de cómputo para el diagnóstico de la probabilidad de un infarto.

Los siguientes pasos esperados son los siguientes:

1. Implementar esta solución para poder crear una dataset cada vez mayor y que pueda incrementarse su precisión.
2. Que el algoritmo sea aplicado en una aplicación o un servicio web para su consumo de cualquier persona u organización clínica.

***10. Bibliografía En***

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304-310.

http://rexa.info/paper/b884ce2f4aff7ed95ce7bfa7adabaef46b88c60c

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

http://rexa.info/paper/0519d1408b992b21964af4bfe97675987c0caefc

Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.

http://rexa.info/paper/faecfadbd4a49f6705e0d3904d6770171b05041f