

1. Descripción general

El objetivo de este proyecto es desarrollar un producto de analítica de datos para la predicción de enfermedades cardíacas. Se han identificado dos usuarios finales interesados en este producto:

1. Médicos que quieran incorporar un apoyo de analítica de datos en el proceso de evaluación de pacientes y la toma de decisiones asociada (solicitud de exámenes, chequeos y otros procedimientos).
2. Pacientes que quieran realizar una prueba sencilla en casa y así obtener información adicional que les acompañe en el proceso de decisión de visitar a un profesional médico.

Seleccione uno de los dos usuarios finales y diseñe su producto pensando en ese usuario. El nivel esperado de desarrollo de este producto es de **prototipo funcional**.

2. Datos

Para desarrollar este producto usted puede hacer uso de los datos disponibles en el repositorio de la Universidad de California en Irvine: <https://archive-beta.ics.uci.edu/ml/datasets/heart+disease>. Allí encontrará datos provenientes de cuatro fuentes, con un número mediano de variables posibles. Note que las variables que se han usado en publicaciones de investigación son 14 del conjunto recopilado en Cleveland. Si prefiere puede utilizar otro conjunto, consúltelo con el instructor.

Tarea 1

Su primera tarea consiste en explorar los datos. Descargue los datos y realice un análisis exploratorio que permita describir estadística y visualmente el comportamiento de las variables a considerar. Calcule estadísticas descriptivas, determine si hay valores faltantes, realice histogramas, diagramas de caja, diagramas de dispersión, diagramas de violín y otros que permitan comprender cómo se comportan las variables. Aquí posiblemente se enfrente a problemas de calidad de datos, como datos faltantes o datos en formatos que no se adecúan a un uso inmediato. Realice y documente los ajustes necesarios.

3. Modelos

Tras explorar en detenimiento los datos, es hora de pasar a construir los modelos. Se espera que sean modelos basados de redes Bayesianas. Tenga presente lo aprendido en la exploración de datos, así como el usuario final seleccionado y sus necesidades.

Tarea 2

Aquí deberá explorar diferentes configuraciones de modelo, emplear diferentes métodos de estimación, comparar y seleccionar las mejores alternativas. Consulte bibliografía que

le permita **proponer una estructura de grafo**, sobre la cual debe emplear métodos de estimación de máxima verosimilitud y Bayesiana para parametrizar su modelo. Para este ejercicio es de esperar que el foco esté en redes Bayesianas con variables discretas, luego es posible que deba discretizar algunas variables para emplearlas en los modelos. No es necesario emplear todas las variables disponibles, pero todas las variables incluidas y sus relaciones deben estar correctamente justificadas.

4. Producto

Tras explorar los datos y construir los modelos, es hora de diseñar y desarrollar el producto. El producto debe ser un tablero en Dash desplegado en la nube, usando una máquina virtual. El tablero debe ser de fácil uso y le debe permitir al usuario ingresar datos para realizar predicciones del riesgo de sufrir una enfermedad cardíaca.

Tarea 3

Empiece por diseñar el tablero: ¿qué valores debe permitir ingresar? ¿qué resultados genera? ¿qué visualizaciones incluye? ¿cómo mostrará las instrucciones? ¿cómo dispondrá estos elementos en el tablero? Para esta tarea es buena idea hacer un wireframe, que es un diseño sencillo (como de servilleta, pero digital) que le permite tener una visión clara de su tablero y todos sus elementos. Recuerde no perder de vista al usuario y su necesidad. Una vez haya terminado el diseño, desarrolle su tablero en Dash y despléguelo en un ambiente de pruebas.

5. Entregables

Como resultado de las tareas anteriores deberá entregar los siguientes resultados y soportes:

1. **(25 puntos)** Resultado 1: reporte de entre 5 y 15 páginas con los resultados principales del análisis exploratorio de datos y la modelización.
2. **(25 puntos)** Resultado 2: presentación de no más de 10 minutos con los resultados principales del análisis exploratorio de datos y la modelización. Esta presentación debe incluir también un espacio para demostrar el tablero desarrollado.
3. **(50 puntos)** Resultado 3: tablero desarrollado en Dash y desplegado en la nube en un ambiente de pruebas.
4. **(5 puntos)** Reporte de trabajo en equipo: incluya un pequeño reporte de cómo se dividieron las tareas entre los miembros del equipo.
5. Soporte 1: fuentes de análisis (cuadernos de jupyter o archivos .py con el análisis exploratorio).

6. Soporte 2: fuentes de análisis (cuadernos de jupyter o archivos .py con la modelización desarrollada). Aquí debe incluir un paso de construcción de la estructura del modelo, un paso de parametrización, y un paso de inferencia.
7. Soporte 3: fuentes del tablero (archivos .py del tablero desarrollado).
8. Soporte 4: repositorio Git en Github, con un historial de commits que claramente refleje el aporte de cada miembro del grupo.

Nota: si bien el trabajo es en grupo (de 2 o 3 personas), la nota es individual, luego es necesario que cada miembro del equipo demuestre su contribución al proyecto, tanto a través de los commits en el repositorio, como a través del reporte de trabajo en equipo y la sustentación.

6. Recomendaciones

1. El objetivo del proyecto es lograr un buen producto, bien soportado y claramente desarrollado. Justifique adecuadamente sus decisiones, observaciones y conclusiones.
2. Sea conciso y eficiente con el espacio. Ni el reporte ni la presentación deben ser largos. Al contrario, en un buen reporte cada gráfica y afirmación importa, y en una buena presentación cada diapositiva cuenta.
3. Es un trabajo en equipo. Defina los ítems de trabajo, asígneles entre los miembros del equipo, defina fechas de entrega y revisión interna. Discuta los resultados, observaciones y conclusiones. Priorice tareas y resultados a incluir.
4. Empiece a trabajar prontamente y discuta con el instructor su avance y resultados.

Fecha de entrega: jueves 9 de marzo