**Índice**

Contenido

[Índice 1](#_heading=h.gjdgxs)

[Puntos a tener en cuenta para TFM 1](#_heading=h.30j0zll)

[Introducción 1](#_heading=h.1fob9te)

[Datos 2](#_heading=h.3znysh7)

[Enlaces de interés 2](#_heading=h.2et92p0)

[Apéndice 3](#_heading=h.tyjcwt)

**Puntos a tener en cuenta para TFM**

1. Replicar su modelo (CancerA1DE), entrenarlo y validarlo con parámetros que ofrezcan resultados parejos haciendo uso de los datos que usaron ellos.
2. Probar el modelo CancerA1DE con datos de fuentes externas, para comprobar si de verdad ofrece una buena predicción.
3. Construir otros modelos y comparar su rendimiento con el CancerA1DE
4. Si da tiempo, profundizar en la detección del tipo de cáncer mediante el modelo CancerA2DE añadiendo a la muestra las 31 variables. Al tener más variables en juego, puede ser más interesante ya que se puede acercar más a un caso de la vida real.
   * Replicar su modelo
   * Construir otros modelos y ver cómo rinden en comparación
   * Si diese tiempo, encontrar data sets externos para poder testearlo

**Introducción**

En este estudio se lleva a cabo el contraste del modelo CancerA1DE para la detección prematura de cáncer frente a otros modelos de predicción, evaluando su eficacia frente a los datos utilizados en este estudio y frente a otros externos. Este estudio de CancerA1DE, se basa a su vez en uno llamado CancerSeek. Los datos provienen de este estudio primigenio.

**Datos**

Los datos provenientes del estudio primigenio se encuentran repartidos en once tablas con información relativa a los pacientes, a las proteínas y marcadores presentes en las analíticas de sangre influyentes para la predicción e información de los resultados recopilada en este estudio original de CancerSeek.

La información relativa a cada una de las tablas se dejará en el anexo.

El primer conjunto de datos tiene 1.817 registros de análisis de sangre de pacientes, que están diseñados y adoptados para construir modelos para detectar cánceres como detector de primera línea de forma binaria (es decir, cáncer o normal). Por lo tanto, para ser escalable y económico, tiene la cantidad mínima de información de características de entrada que involucra ocho concentraciones de marcadores de proteínas circulantes y una puntuación de mutación de ADN libre de células (OmegaScore).

El segundo conjunto de datos tiene 626 registros de análisis de sangre de pacientes, que están diseñados y adoptados para construir modelos para localizar tipos de cáncer como diagnóstico de segunda línea (es decir, mama, colorrectal, tracto gastrointestinal superior, hígado, pulmón, ovario o páncreas). Por lo tanto, su conjunto de funciones de entrada cubre las nueve funciones anteriores e incluye 31 marcadores de proteínas adicionales y el sexo del paciente.

Todos los datos de las proteínas y biomarcadores de estos registros se encuentran en la tabla S6. De esta se extraen las muestras de los 8 biomarcadores de mayor peso para el predictor de primera línea. Luego de esa misma tabla, se toman los registros correspondientes para el detector del tipo de cáncer, cuya información se guarda en la tabla S8.

Durante el análisis de los datos se observa que muchas de las muestras de los parámetros de las analíticas, que normalmente se almacenarían como numéricas, se encuentran en formato categórico. Este enfoque resulta del estudio CancerA1DE, donde se aplica el principio MDL (Minimum Description Length), basado en la idea de que el mejor modelo es aquel que minimiza la longitud de la descripción del conjunto de datos, teniendo en cuenta tanto la capacidad de ajuste del modelo como la simplicidad de este.

Por tanto, trata de encontrar la mejor manera de dividir los valores de una característica continua en intervalos discretos o categorías.

Preprocesado de los datos

Enlaces de interés

* Artículo principal CancerA1DE & CancerA2DE: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548890/>
* Artículo del estudio original CancerSEEK : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080308/>
* Técnica “Not So Naive Bayes” utilizada en el artículo principal : <https://link.springer.com/content/pdf/10.1007/s10994-005-4258-6.pdf>
* Paper con info sobre la utilización algoritmo Bayesiano generativo AnDE y cómo rinde frente a otras técnicas : <https://link.springer.com/content/pdf/10.1007/s10994-011-5263-6.pdf>
* Modelos Deep Learning. Algunos utilizados en el estudio principal : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965871/>
* Base de datos STRING donde se puede sacar relaciones entre las proteínas analizadas en las analíticas y que usan en el artículo principal [complejo] : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210637/>

**Información extra**

* Web de la Agencia Internacional de estudio del cáncer. Figuras, tablas e imágenes de la situación del cáncer en el mundo. [chulo] : <https://gco.iarc.fr/today/en>
* Situación actual del cáncer en EEUU : <https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.3322/caac.21820>
* Estadísticas de la situación del cáncer en 2015 en China : <https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.3322/caac.21338>
* Estadísticas globales del cáncer en 2012 : <https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.3322/caac.21262>

**Anexo**

Información de las tablas

Tablas S1 a S11:

Tabla S1 (Formato Microsoft Excel): Secuencias de cebadores para ensayos de PCR múltiple.

Tabla S2 (Formato Microsoft Excel): Mutaciones identificadas en tumores primarios.

Tabla S3 (Formato Microsoft Excel): Biomarcadores proteicos analizados e incluidos en la prueba CancerSEEK.

Tabla S4 (Formato Microsoft Excel): Características histopatológicas y clínicas de los pacientes con cáncer y los controles sanos.

Tabla S5 (Formato Microsoft Excel): Mutaciones identificadas en muestras de plasma de pacientes con cáncer y controles sanos.

Tabla S6 (Formato Microsoft Excel): Concentraciones de biomarcadores proteicos ensayados en muestras de plasma de pacientes con cáncer y controles sanos.

Tabla S7 (Formato Microsoft Excel): Concordancia entre las mutaciones identificadas en el plasma y aquellas identificadas en tumores primarios.

Tabla S8 (Formato Microsoft Excel): Resultados de localización del tipo de cáncer para los 617 pacientes con cáncer identificados por CancerSEEK.

Tabla S9 (Formato Microsoft Excel): Coeficientes y puntajes de importancia del modelo de regresión logística.

Tabla S10 (Formato Microsoft Excel): Matriz de confusión de las principales predicciones de los resultados de localización del tipo de cáncer.

Tabla S11 (Formato Microsoft Excel): Pacientes con cáncer evaluados en este estudio por tipo y etapa de tumor.