Proyecto de Ciencia de Datos

Grado en Ciencia de Datos e Inteligencia Artificial

Segundo proyecto

Enunciado

Este proyecto trata sobre la ejecución técnica de un proyecto de ciencia de datos. Para ello se cuenta con datos (sintéticos) de pacientes con cáncer de mama que hemos recibido en diferentes formatos (CSV, Excel, JSON y como dump de una base de datos).

El trabajo consiste en formular un objetivo de negocio con algún indicador de rendimiento y su transformación en objetivo de ciencia de datos. En base a ellos, se debe realizar todo el pre-procesamiento que se considere necesario sobre los datos tras la exploración de los mismos. Finalmente es necesario realizar los análisis descriptivos y/o predictivos que se deseen.

Para formular estos objetivos, se es necesario considerar los costes del tratamiento. Para ello podemos saber que actualmente se dispone de un tratamiento novedoso de alto coste que es capaz de reducir la mortalidad un 60%, pero tiene unos graves efectos secundarios sobre las pacientes, por lo que sólo debe administrarse en casos con mal pronóstico. El coste de este tratamiento es de 150.000€ por paciente, frente al coste habitual de una paciente que se somete a mastectomía y quimioterapia convencional (que puede estimare a través de fuentes públicas). Además, para las pacientes con alto riesgo de recaída se puede hacer un seguimiento especial con un coste de 40.000€, que puede reducir las recaídas en un 80% ya que, en esos casos sólo se realizará una mastectomía extra sin que el cáncer llegue a desarrollarse.

Realización y evaluación

El proyecto se realizará en grupos de 4 estudiantes y se evaluará mediante una presentación de 15 minutos del mismo. En la presentación es necesario indicar los objetivos planteados, los pasos más importantes realizados en el pre-procesamiento, el análisis descriptivo y/o predictivo y su relación con los objetivos a través de los costes establecidos.

La presentación se realizará el día del examen ordinario de la asignatura, en el aula y hora establecidas por jefatura de estudios (jueves, 25 de enero de 2024 a las 15:00).

Se habilitará una entrega en Moodle para entregar la presentación que se realizará (no es necesario entregar reporte, sólo la presentación) y el código usado para el desarrollo. La tarea estará disponible para su entrega hasta el día de antes de la presentación a las 23:59, es decir, hasta el 24 de enero de 2024.

En la calificación de este proyecto también se considerarán los ejercicios realizados y entregados en clase.

Variables proporcionadas

Además de las variables mencionadas en cada apartado, siempre se incluye el ehr, que es un identificador de la paciente.

- patients (dividido en los batches de datos recibidos en momentos diferentes):
  - birth_date: fecha de nacimiento de la paciente.
  - diagnosis_date: fecha de diagnóstico del primer tumor.
  - death_date: fecha de fallecimiento (un nulo indica que la paciente no ha fallecido).
- gynecological:
  - pregnancy: número de embarazos de la paciente.
  - birth: número de partos naturales de la paciente.
  - caesarean: número de partos por cesárea de la paciente.
  - abort: número de abortos de la paciente.
  - menarche_age: edad que la paciente tenía en su primera menstruación.
  - menopause_age: edad que la paciente tenía al entrar en la menopausia.
- tumor:
  - n_tumor: identificador del tumor, correlativo para cada paciente.
  - t_category: indicador del tamaño del tumor. Valores posibles: IS, 0, 1, 2, 3, 4.
  - n_category: indicador del número de ganglios linfáticos afectados por el tumor. Valores posibles: 0, 1, 2, 3.
  - m_category: indicador de si hay metástasis a distancia. Valores posibles: 0, 1.
  - stage_diagnosis: estadio del tumor al diagnóstico (puede derivarse del TNM). Valores posibles: 0, IA, IB, IIA, IIB, IIIA, IIIB, IIIC, IV.
  - t_category_after_neoadj, n_category_after_neoadj, m_category_after_neoadj y stage_after_neo: mismo significado que las anteriores, pero medidas tras el tratamiento neoadyuvante (si lo hubo).
  - grade: grado del tumor. Valores posibles: 1, 2, 3.
  - ductal: indica si el tumor es de tipo histológico ductal. Valores posibles: 0, 1.
  - lobular: indica si el tumor es de tipo histológico lobulillar. Valores posibles: 0, 1.
  - neoadjuvant: indica si la paciente recibió tratamiento neoadyuvante. Valores posibles: 0, 1.
- histochemistry:
  - er: indica si la paciente tiene receptores de estrógenos. Valores posibles: 0, 1.
  - pr: indica si la paciente tiene receptores de progesterona. Valores posibles: 0, 1.
  - her2: indica si la paciente tiene el gen HER2 sobreexpresado. Valores posibles: 0, 1.
  - ki67: índice de proliferación del tumor (%).