CURSO DE APRENDIZAJE AUTOMÁTICO

TAREA FINAL

Contexto

El seguro de vida o seguro sobre la vida es un seguro que cubre el riesgo de muerte o incapacidad. Este seguro es formalizado mediante un contrato o póliza entre la empresa aseguradora y el asegurado.
La empresa aseguradora garantizará una indemnización en caso de fallecimiento o invalidez del asegurado. El cálculo del costo del seguro se verá afectado por la edad del asegurado y el capital a asegurar. Al mismo tiempo, existen otras variables que afectarán el costo del seguro o inclusive que pueden llevar a ser negado por parte de la empresa aseguradora, entre ellas, la práctica de deportes de algo riesgo, una profesión riesgosa o enfermedades prexistentes. Dentro de las enfermedades prexistentes, las afecciones cardíacas son las que tienen mayor peso en la decisión por parte de la empresa aseguradora.

Descripción del problema

Una empresa aseguradora ha detectado que en los últimos años ha negado una gran cantidad de seguros de vida asociados a hipotecas. Todos estos casos se han debido a que, en el cuestionario de alta del seguro, el asegurado respondió que ha sufrido de alguna afección cardíaca. Sin embargo, un análisis posterior, permitió detectar que este criterio ha sido mal aplicado, en la mayoría de los casos se podría haber entregado el seguro. Estas malas decisiones han hecho perder clientes, y, por consiguiente, ingresos.
La empresa ha tomado la decisión de solicitar análisis clínicos a aquellos potenciales asegurados que en el cuestionario de alta se desprenda riesgo de afección cardíaca. El objetivo es utilizar los resultados clínicos para inferir si el cliente tiene un alto riesgo de muerte, solo en este caso se negará la póliza.

Actualmente, los datos que son solicitados al potencial asegurado son los siguientes:

- Nombres y apellidos
- Fecha de nacimiento
- Peso y estatura
- Deportes que practica
- ¿Actualmente fuma o ha fumado?
- ¿Realiza alguna actividad de riesgo?

- ¿Fue informado alguna vez que presentaba cifras elevadas de tensión arterial o le fue prescripto algún tratamiento para la hipertensión arterial?
- ¿Padece o padeció enfermedades cardiovasculares (infarto, angina de pecho, arritmia, cardiopatía, etc.)?
- ¿Recibe actualmente o recibió alguna vez tratamiento a causa de diabetes, colesterol, triglicéridos, hormonales, gota, cáncer o tumores?

La empresa aseguradora se encuentra en un fuerte proceso de transformación digital, por lo que pretende utilizar técnicas de aprendizaje automático para la toma de decisión del otorgamiento de la póliza. Para esto, ha logrado obtener un dataset con valiosos datos, a partir de la investigación de *Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5*

Parte 1

Efectúe un profundo EDA del dataset entregado. En caso de que existan dudas de la semántica de alguno de los features del dataset, se recomienda fuertemente investigar sobre el mismo y documentar el resultado.

Parte 2

Utilizando las técnicas vistas en el curso, detecte las características de las poblaciones con mayor riesgo de muerte producto de un accidente cardiovascular

Parte 3

Para las poblaciones de mayor riesgo, en función de las respuestas obtenidas en el formulario de alta, determine el conjunto de resultados clínicos que el potencial asegurado debe entregar para avanzar en el proceso de obtención de la póliza.

Parte 4

Para el conjunto de respuestas obtenidas en el formulario de alta, diseñe un modelo predictivo que determine si el potencial asegurado requiere de la realización de análisis clínicos adicionales o la póliza le será entregada sin averiguaciones adicionales.
Documente detalladamente todas las decisiones tomadas.
Actualmente no existen formularios digitalizados con las respuestas de los asegurados, por lo que es necesario la generación de datos sintéticos que representen las distintas poblaciones. Esto datos serán los utilizados para el entrenamiento y test del modelo predictivo.

Consideraciones generales

- Como ambiente de trabajo puede utilizar Jupyter Notebooks o Google Colab
- Si usa Jupyter Notebooks, el lenguaje de programación seleccionado debe ser Python
- Se recomienda fuertemente el uso del paquete scikit-learn
- Cada grupo debe entregar
    - Notebook
    - En caso de que documente fuera del notebook, entregar el documento anexo en formato PDF
    - Siempre incluya número de grupo, junto con el nombre de cada uno de los estudiante que lo integran
- **El plazo de entrega vence el domingo 14 de noviembre a las 23:59**. La entrega se efectuará vía Moodle. **Las defensas se llevarán a cabo el martes 16 y el jueves 18 de noviembre.**