INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

Tarea 3

Predicción de etapa de cirrosis

Profesor: Juan D. Velásquez
Auxiliares: José I. Saffie y José F. Soza

Descripción del problema

Usted y su talentoso equipo de Machine Learning demostraron poder hacer una correcta clasificación de los hongos que atacaron a la sociedad de los túneles, por lo que jamás los iban a dejar ir tan fácil sin aprovechar sus últimos conocimientos en aprendizajes supervisados como no supervisados.

Como último requerimiento del emperador, este le comunica que últimamente la civilización ha estado padeciendo una enfermedad que deteriora el hígado de forma progresiva, la cual es la temida cirrosis. Ellos desean poder saber en qué estado está esta enfermedad en uno de los habitantes, con tal de poder tomar mejores decisiones dependiendo de cuánto ha avanzado la cirrosis, sin embargo, los tests para evaluar esto se tardan demasiado y son muy caros, por lo que les gustaría poder determinar de una forma más rápida el estado de la enfermedad y sabe que usted y su equipo pueden hacerlo.

Para ello, la civilización cuenta con 25.000 registros y distintas características de personas que padecen o padecieron esta enfermedad. Su detalle se encuentra en el diccionario al final del enunciado.

Preguntas

P0.- Realice un análisis exploratorio de los datos (EDA) y aplique la metodología de Feature Engineering. No es necesario generar nuevas variables (Feature Generation), pero si debe descartar aquellas variables que no aporten a la construcción del modelo. **(1 pto.)**

P1.- Construya un modelo de clasificación con cada una de las siguientes técnicas para el problema de clasificar el estado de cirrosis de una persona: Decision Trees, Support Vector Machines, KNN y dos métodos ensamblados a su elección. Además, asegúrese de realizar un ajuste de los hiperparámetros de los modelos. Finalmente, indique qué modelo (y con qué hiperparámetros) es mejor a partir de una evaluación realizada con las métricas vistas en clases, matriz de confusión y área bajo la curva ROC (con su respectivo gráfico). **(2.5 ptos.)**

P2.- Suponga que la civilización desea hacer una segmentación de sus habitantes en base a diferentes características de estos, con tal de detectar nuevos grupos de personas (y no solo por estado de cirrosis), no obstante, desconocen cómo hacerlo. Para esto, realice un proceso de clustering particional, clustering jerárquico (con su respectivo dendrograma) y clustering basado en densidad.

Indique claramente qué variables utilizará para clusterizar, qué cantidad de clusters utilizará y por qué. Finalmente, evalúe los clusters construidos a través del silhouette score. **(2.5 ptos.)**

BONUS.- Realice un Análisis de Componentes Principales para poder graficar los clusters en 2 dimensiones. Comente los resultados. ¿La medida fue adecuada o no? **(0.5 ptos.)**

Reglas del juego

- Debe realizar esta tarea con su grupo de trabajo, publicado a través de U-Cursos.
- La resolución de dudas vía canales virtuales será realizada de forma exclusiva a través del foro del curso.
- Debe trabajar exclusivamente utilizando lenguaje Python en un documento *notebook* de *Google Colab*. Para ello, genere una copia de la plantilla del documento de Colab disponible en este link. Puede agregar las celdas de código y texto que estime conveniente, siempre y cuando mantenga la estructura de secciones previamente definida en la plantilla.
- Las conclusiones preliminares y finales, así como su trabajo en general, deben ser congruentes con el objetivo del problema de construir un modelo de clasificación y clusterizaciones.
- El formato de su entrega debe ser el de un informe autocontenido, es decir, debe contener una introducción al trabajo a realizar, además de conclusiones sobre sus resultados. Cada etapa de su procedimiento debe contener celdas de texto explicando qué está realizando. Para ello, aproveche el paradigma de *programación literaria* de Google Colab que le permite intercalar bloques de texto y código. Al respecto, **código y procedimientos no explicados serán evaluados con puntaje cero.**
- Puede usar los ejemplos de código vistos en las clases auxiliares.
- Su entrega debe consistir de el documento notebook `.ipynb`, además de un archivo `.txt` que contenga el link hacia su notebook en Google Colab, **en modo lector**.
- Fecha de entrega: Lunes 1 de julio a las 23:59

Anexo: Diccionario

1. **N_Days:** Número de días entre el registro y el evento más temprano entre muerte, trasplante o el momento del análisis del estudio en 1986.
2. **Status:** Estado del paciente: C (censurado), CL (censurado por trasplante de hígado) o D (muerte).
3. **Drug:** Tipo de droga: D-penicilamina o placebo.
4. **Age:** Edad en días.
5. **Sex:** Sexo: M (masculino) o F (femenino).
6. **Country:** País.
7. **Ascites:** Presencia de ascitis: N (No) o Y (Sí).
8. **Hepatomegaly:** Presencia de hepatomegalia: N (No) o Y (Sí).
9. **Spiders:** Presencia de arañas vasculares: N (No) o Y (Sí).
10. **Edema:** Presencia de edema: N (sin edema y sin terapia diurética para el edema), S (edema presente sin diuréticos o edema resuelto por diuréticos) o Y (edema a pesar de la terapia diurética).
11. **Bilirubin:** Bilirrubina sérica en [mg/dl].
12. **Cholesterol:** Colesterol sérico en [mg/dl].
13. **Albumin:** Albúmina en [gm/dl].
14. **Sodium:** Niveles de sodio en [mEq/L].
15. **Copper:** Cobre en orina en [ug/día].
16. **Alk_Phos:** Fosfatasa alcalina en [U/litro].
17. **SGOT:** SGOT (enzima) en [U/ml].
18. **Tryglicerides:** Triglicéridos en [mg/dl].
19. **Platelets:** Plaquetas por [ml/1000].
20. **Prothrombin:** Tiempo de protrombina en segundos [s].
21. **Stage:** Etapa histológica de la enfermedad (1, 2 o 3).