

## **Proyecto 2 de Minería de Datos**

### **Predicción de muerte por falla cardiovascular**

Las enfermedades relacionadas con el corazón o las enfermedades cardiovasculares (ECV) son la razón principal de una gran cantidad de muertes en el mundo en las últimas décadas y se han convertido en la enfermedad que más amenaza la vida, no solo en la Chile sino que en todo el mundo. Por lo tanto, existe la necesidad de un desarrollar un sistema confiable, preciso y factible para detectar tales enfermedades a tiempo para así lograr intervenir y establecer un tratamiento oportuno y adecuado a cada paciente.

Día a día, los casos de éste tipo de enfermedades están aumentando a un ritmo acelerado. Es por esta razón que predecir posibles muertes de dichas enfermedades ayuda al proceso de diagnóstico en conjunto con un análisis completo del estilo de vida del paciente a evaluar.

La mayoría de las enfermedades cardiovasculares se pueden prevenir considerando factores de riesgo conductuales, como el tabaquismo, la dieta poco saludable y la obesidad, la inactividad física y el consumo nocivo de alcohol, mediante estrategias que abarquen a toda la población. Las personas con enfermedades cardiovasculares o que tienen un alto riesgo cardiovascular (por la presencia de uno o más factores de riesgo como hipertensión, diabetes, hiperlipidemia o enfermedad ya establecida) y necesitan una detección y manejo temprano , situación que bajo un enfoque de clasificación un modelo de aprendizaje automático puede ser de gran ayuda.

El objetivo de este proyecto es generar un programa computacional basado en aprendizaje supervisado que logre predecir la presencia de una enfermedad cardíaca de un individuo (El campo "Target" se refiere a la presencia de enfermedad cardiaca en el paciente. Tiene un valor entero 0 = sin enfermedad y 1 = enfermedad). Para dicha misión usted cuenta con un conjunto de datos que contiene 14 atributos (heart.csv).

Como resultado del trabajo, genere un archivo jupyter (.ipynb) mostrando su proceso. El archivo deberá tener un reporte describiendo las etapas de su proceso (limpieza de datos, selección de variables, y otras), a lo menos 3 modelos supervisado, la selección de hiperparámetros (justificada empíricamente), la comparación de los resultados obtenidos entre los modelos, y el código empleado.

Los puntajes asignados a cada tarea corresponden a:

- Limpieza de datos (1.0 punto, código y justificación)
- Modelo de clasificación, 3 modelos (3.0 puntos, código y justificación de los hiperparámetros de cada modelo)
- Selección del modelo final (2.0 puntos, código y justificación)

**El trabajo es en grupos de máximo 3 alumnos.**

**El resultado de su proyecto deberá ser subido a webcursos a más tardar el 26 de junio a las 23:59 horas.** Usted podrá subir los archivos requeridos o subir los códigos a su cuenta de GitHub de alguno de los participantes y enviar el enlace a través de WebCursos.

**¡ Mucha Suerte!**

HINT: Variables

1. Age = edad
2. Sex = sexo/genero
3. chest pain type (4 values) = tipo de dolor torácico
4. resting blood pressure = presión arterial en reposo
5. serum cholestoral in mg/dl = colesterol sérico
6. fasting blood sugar > 120 mg/dl = Azúcar en sangre en ayunas
7. resting electrocardiographic results (values 0,1,2) = resultados electrocardiográficos en reposo
8. maximum heart rate achieved = frecuencia cardíaca máxima alcanzada
9. exercise induced angina = angina inducida por el ejercicio
10. oldpeak = ST depression induced by exercise relative to rest = depresión del ST inducida por el ejercicio en relación con el reposo
11. the slope of the peak exercise ST segment = la pendiente del segmento ST de ejercicio máximo
12. number of major vessels (0-3) colored by flourosopy = número de vasos principales (0-3) coloreados por fluoroscopia

13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values. tal: 0 = normal; 1 = defecto fijo; 2 = defecto reversible