Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (II/2021)

Actividad 4

Objetivos

- Aplicar los contenidos de análisis de datos con Python para procesar y visualizar información, y realizar predicciones.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **A4**.
- **Entrega: lunes 15 de noviembre a las 18:30 hrs.**
- **Formato de entrega:** archivo Python Notebook (**A4.ipynb**) y archivo Python (**A4.py**) con la solución de este enunciado. Los archivos deben estar ubicados en la carpeta **A4**. No se debe subir ningún otro archivo a la carpeta. Utilice múltiples celdas de texto y código para facilitar la revisión de su programa.
- **NO SE ADMITEN ENTREGAS FUERA DE PLAZO**
- **Entregas con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**

Introducción

Con el fin de evaluar los contenidos de análisis de datos con Python, en esta actividad deberá realizar procesamiento de datos para finalmente predecir la existencia de problemas cardíacos en pacientes.

Descripción de los datos

La fuente primaria de datos para será un subconjunto del set "Heart Disease", disponible en el sitio del curso. El set contiene información sobre la existencia de problemas cardíacos en pacientes, donde además se indica la severidad de la patología en caso de que esta exista. Específicamente, el archivo `heart.csv` contiene registros de 303 pacientes, donde cada uno incluye 13 características y 1 variable a predecir. La descripción de cada una de estas se entrega a continuación:

- age: edad del paciente.
- sex: sexo del paciente.
- cp: tipo de dolor de pecho.
- trestbps: presión sanguínea en reposo.
- chol: colesterol.
- fbs: nivel de azucar en la sangre (glicemia).
- restecg: resultados electrocardiográficos en reposo.
- mhr: máxima frecuencia cardíaca alcanzada.
- exang: dolor producido por ejercicio.
- oldpeak: depresión inducida por ejercicio en segmento ST de un electrocardiograma.
- slope: pendiente del electrocardiograma en el segmento ST al nivel máximo de ejercicio.
- ca: número de vasos sanguíneos mayores coloreado por la fluoroscopía.
- thal: nivel de Talasemia.
- target: existencia o nivel de gravedad de la enfermedad cardíaca (variable a predecir).

IMPORTANTE

Para cumplir las misiones de esta actividad, es su responsabilidad explorar inicialmente el contenido de los archivos y familiarizarse con el formato en que está almacenada la información.
Recuerde además codificar numéricamente los valores de las columnas categóricas y normalizar las numéricas, cuando corresponda. La actividad no considera puntaje por hacer esto, pero sí descuentos cuando no es realizado o es realizado en una variable o momento incorrecto.

Misión 1: completando información

a) Identifique qué columna(s) deben ser preprocesadas y corríjalas, utilizando algún criterio que respete la distribución de los datos de la(s) columna(s). Recuerde que la variable a predecir no debe ser preprocesada. **(1 pto.**)

b) Identifique las columnas que presentan correlación fuerte (≥ 0.7), sin considerar en el análisis a la variable a predecir. Luego, por cada par identificado, descarte una de las dos columnas. **(1 pto.**)

Misión 2: predicción de la patología

Construya modelos que permitan predecir la existencia y severidad de la patología, dadas las características del paciente. En particular, deberá evaluar diversas estrategias para construir modelos, reportando en cada caso el rendimiento en un conjunto de prueba independiente de los datos usados para el entrenamiento.

a) Predicción binaria: entrene 3 modelos que permitan predecir la existencia o no de patología cardíaca, sin considerar la severidad de estas (es decir, se debe predecir SÍ o NO). **(1 pto.**)

b) Predicción multiclase: entrene 3 modelos que permitan predecir la existencia y severidad (en caso de existir) de patologías cardíacas (es decir, se debe predecir lo que indica la columna `target`). **(1 pto.**)

c) Predicción binaria jerárquica: entrene un modelo compuesto que realice la misma predicción del ítem anterior, pero de manera jerárquica y binaria, es decir, discriminando entre una categoría y el resto de ellas de manera iterativa, hasta cubrir todos los casos. Por ejemplo, el primer submodelo puede indicar si un paciente tiene una patología o no. Luego, el siguiente submodelo de la jerarquía puede indicar, para los pacientes con enfermedad cardíaca, si es de nivel 1 o mayor. El siguiente, si es 2 o mayor, y así sucesivamente hasta cubrir todos los casos. El orden en que se analizan las categorías y el clasificador utilizado para cada submodelo es una decisión que debe tomar ud. **(2 ptos.**)

Política de Integridad Académica

"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad."

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por "copia" o "plagio" se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.