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

IIC2115 – Programación como Herramienta para la Ingeniería (II/2023)

Tarea 1

Aspectos generales

- **Objetivo:** aplicar los contenidos de análisis y visualización, y modelos predictivos de aprendizaje de máquina para estudiar, visualizar y predecir propiedades o relaciones que se pueden observar en en conjunto de datos de transporte público.
- **Lugar de entrega:** domingo 15 de octubre hasta las 23:59 hrs. en repositorio privado.
- **Formato de entrega: ÚNICAMENTE** el archivo Python Notebook (**T1.ipynb**) con la solución de la tarea. El archivo debe estar ubicado en la carpeta **T1**. Es requerimiento de formato el utilizar múltiples celdas de texto y código para la construcción de la solución, además de dejar todas estas ejecutadas en el archivo entregado. Entregas que no cumplan el formato de entrega tendrán un descuento de 0,5 ptos.
- **Entregas atrasadas:** El descuento por atraso se realizará de acuerdo a lo definido en el programa del curso.
- **Issues:** las discusiones en las *issues* del Syllabus que sean relevantes para el desarrollo de la tarea, serán destacadas y se considerarán como parte de este enunciado.
- **Librerías:** el uso de librerías externas, distintas a las cubiertas en el curso y que solucionen aspectos centrales o extensos de la tarea está prohibido. Solo se podrán usar las que han sido aprobadas explícitamente por el cuerpo docente, previa consulta de los estudiantes, lo que será informado en el Syllabus a través de *issues*.
- **Tarea con errores de sintaxis y/o que generen excepciones en todas las ejecuciones serán calificados con nota 1,0.**

Introducción

Estudio de datos de buses del transporte público metropolitano

Usted ha sido contratado por el Directorio de Transporte Público Metropolitano (DTPM) para analizar el rendimiento de los buses del transporte público. DTPM es el organismo público encargado de analizar de forma integral el sistema de transporte público capitalino y velar por la adecuada coordinación de los diferentes modos que de este. Es por esto que está muy interesado en extraer conocimiento desde los datos generados diariamente por los dispositivos GPS presentes en cada bus.

El conjunto de datos

Los datos se encuentran almacenados en un archivo de texto plano de valores separados por coma (**.csv** del inglés *comma-separated values*) llamado `mediciones_gps.csv`. Cada fila del archivo posee valores separados por una coma (**","**). Los valores de la primera fila corresponden a los nombres de las columnas y el resto de las filas contienen la información asociada a las mediciones de los GPS de los buses. Cada fila corresponde a 1 medición de 1 GPS de 1 bus en 1 instante determinado y las columnas (o atributos) que posee un pulso se describen a continuación:

1. **measurement_id**: identificador interno de una medición GPS. Este valor corresponde a un número hexadecimal de largo fijo.

2. **expedition_id**: identificador interno de la expedición asociada a la medición GPS. Una expedición corresponde a la realización de un recorrido por un bus. Este valor corresponde a un número hexadecimal de largo fijo.

3. **dispatch_time**: instante de tiempo en que ocurre el despacho de la expedición asociada a la medición GPS. Este valor se encuentra representado en el formato `año-mes-dia hora:minuto:segundo`, por ejemplo, "2018-04-01 14:33:21".

4. **line_id**: identificador interno de la línea del bus. La línea de un bus se refiere a servicio prestado por el sistema. Este valor corresponde a un número hexadecimal de largo fijo.

5. **line_code**: código usuario de la línea. Este corresponde al código que llevan los buses en su parte frontal. Se encuentra represando por el código, por ejemplo "C02".

6. **direction**: sentido de operación de la línea. Este campo indica si la operación es en el sentido "ida", representado por una **I** o "regreso", representado por una **R**.

7. **bus_id**: identificador interno del bus. Este valor corresponde a un número hexadecimal de largo fijo.
8. **license_plate**: placa patente del bus. Por ejemplo "BFRD-27" o "FG-3241".
9. **bus_capacity**: capacidad física máxima del bus. Representado con un número que representa la cantidad de personas que caben en el bus.
10. **gps_time**: instante de tiempo en que ocurre la medición GPS. Este valor se encuentra representado en el formato `año-mes-dia hora:minuto:segundo`, por ejemplo, "2018-04-01 14:33:21".
11. **latitude**: latitud de la medición GPS.
12. **longitude**: longitud de la medición GPS.
13. **distance_kms**: distancia recorrida por el bus desde el despacho hasta el instante de la medición GPS en kilómetros.
14. **total_kms**: distancia total a recorrer por el bus en kilómetros.
15. **measurement_speed**: estimación de la velocidad en km/h para el instante de la medición GPS, basada en las mediciones previas del mismo bus.
16. **instant_speed**: velocidad instantánea al momento de registrar la medición GPS en km/h.

Misiones

Desarrolle cada misión en un grupo de celdas separado, es decir, no debe combinar desarrollos de varias misiones en una misma celda. Indique con una celda de texto el inicio de la misión que se trabajará a continuación. Las misiones 1 a la 8 tienen un puntaje de 0,5 ptos. cada una, mientras que las misiones 9 y 10 de 1 pto. cada una.

M1. ¿Cuántas mediciones GPS se realizan por servicio? Utilice una función de `pandas` que le permita responder rápidamente esta pregunta. Con la información obtenida indique el porcentaje del total de mediciones GPS para cada servicio.
Resultado esperado: lista de tuplas con los porcentajes de mediciones (servicio, porcentaje).

M2. Construya un gráfico que permita identificar los servicios que poseen la mayor dispersión de la velocidad instantánea
Resultado esperado: gráfico que permite visualizar lo requerido y análisis de los observado.

M3. Grafique la trayectoria de una expedición. Considere que la fecha-hora del despacho es el **minuto 0**, utilice las columnas `distance_kms` y `gps_time` para graficar una curva x [km] vs t [min] de la trayectoria de la expedición.

Resultado esperado: defina una función que reciba `line_code`, `direction` y `expedition_id`, y que muestre el gráfico de la trayectoria. **BONUS:** en vez de recibir el números de la expedición, reciba una lista de número de expediciones y presente en un mismo gráfico todas las trayectorias de la lista de expediciones.

M4. Determine las variaciones de velocidad en la trayectoria de un bus. Para llevar a cabo esta misión, utilice la misión anterior para construir la trayectoria real de un bus. Luego, construya una trayectoria para un bus ficticio (para el mismo rango de tiempo) a velocidad constante e igual a la velocidad media del bus real. Se recomienda construir esta última curva utilizando una función `lambda`. Presente ambas curvas en un mismo gráfico (la trayectoria real y la trayectoria a velocidad constante).

Resultado esperado: defina una función que reciba `line_code`, `direction` y `expedition_id`, y que muestre el gráfico con ambas curvas.

M5. El objetivo de esta misión será construir un `DataFrame` que posea información relevante para estudiar los tiempos de viaje de las expediciones. Para eso, con la ayuda de la función `loc`, filtre los datos para obtener solo la última medición GPS de cada expedición. Recuerde que estas poseen el tiempo de despacho y el tiempo de la última medición. Genere un `DataFrame` que posea solo las columnas `expedition_id`, `line_code`, `direction`, `license_plate`, `bus_capacity`, `total_kms`, `dispatch_time` y `gps_time`.

Resultado esperado: un nuevo `DataFrame` filtrado, con las columnas solicitadas. Muestre las 5 primeras filas del `DataFrame`.

M6. Cree un nuevo `DataFrame` con información agregada para cada una de las expediciones, considerando las columnas `expedition_id`, `week_day`, `reception_time`, `travel_time` y `total_mean_speed`, que representan: el identificador de la expedición, el día de la semana del despacho ("L", "M", "W", "J", "V", "S" y "D"), la hora de término de una expedición, tiempo total de viaje en minutos y la velocidad media, respectivamente.

Resultado esperado: `DataFrame` con las nuevas columnas. Muestre las 5 primeras filas del `DataFrame`.

M7. Para un servicio en específico, ¿qué día se producen más despachos?. En rangos de dos horas para días laborales ¿qué horario presenta la mayor cantidad de despachos? Utilice gráficos para entregar su respuesta.

Resultado esperado: construya una función que reciba el código del servicio y retorne el día con más despachos y el horario laboral (rango de dos horas) con más despachos. Además muestre los gráficos que justifiquen visualmente la respuesta. **HINT:** Agregue los días/horarios antes de calcular.

M8. Para un servicio en específico, en rangos de dos horas ¿cómo distribuyen los tiempos de viaje en días laborales? ¿Cómo distribuyen los tiempos de viaje en fin de semana? Apóyese de gráficos que permitan justificar su respuesta.
Resultado esperado: construya una función que reciba el código del servicio y muestre las distribuciones de tiempos de viaje. **HINT:** Agregue los horarios por días antes de calcular.

M9. Usando la información generada en las misiones anteriores, construya un modelo de predicción de tiempo de viaje, dado un servicio de bus y un horario. Es libre de definir su modelo y variables a considerar, además de la granularidad del horario y del tiempo a predecir.
Resultado esperado: procedimiento completo de preparación de los datos, además del entrenamiento del modelo predictivo. Evaluación y análisis del rendimiento del modelo y de los resultados. **BONUS:** para cada predicción de tiempo de viaje, estime además su varianza.

M10. Usando la información generada en las misiones anteriores y combinando técnicas de aprendizaje no supervisado, realice una agrupación de los servicios, con el fin de identificar aquellos que en el global presentan un comportamiento similar.
Resultado esperado: procedimiento completo de preparación de los datos, además del entrenamiento del o los modelos no supervisados. Se espera además que se presenten visualizaciones y se realice un análisis de los resultados, indicando algunos de los servicios similares identificados.

Corrección

Es importante que deje todas las celdas de su trabajo ejecutadas antes de subir el archivo, de lo contrario se le aplicará un descuento de 0,5 ptos. al puntaje total.
Para la corrección se revisarán los procedimientos desarrollados para responder las diferentes misiones propuestas y la estructura de cómo utiliza las librerías en ellas. Dado lo abierto de las misiones, se espera que las respuestas incluyan análisis y visualizaciones que permitan justificar las decisiones tomadas.

Política de Integridad Académica

Los/as estudiantes de la Escuela de Ingeniería de la Pontificia Universidad Católica de Chile deben mantener un comportamiento acorde a la Declaración de Principios de la Universidad. En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los/as estudiantes que incurran en este tipo de acciones se exponen a un Procedimiento Sumario. Es responsabilidad de cada estudiante conocer y respetar el documento sobre Integridad Académica publicado por la Dirección de Docencia de la Escuela de Ingeniería.

Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica. Todo trabajo presentado por un/a estudiante para los efectos de la evaluación de un curso debe ser hecho **individualmente** por el/la estudiante, **sin apoyo en material de terceros**. Por "trabajo" se entiende en general las interrogaciones escritas, las tareas de programación u otras, los trabajos de laboratorio, los proyectos, el examen, entre otros.

En particular, si un/a estudiante copia un trabajo, o si a un/a estudiante se le prueba que compró o intentó comprar un trabajo, **obtendrá nota final 1.1 en el curso** y se solicitará a la Dirección de Pregrado de la Escuela de Ingeniería que no le permita retirar el curso de la carga académica semestral.

Por "copia" se entiende incluir en el trabajo presentado como propio, partes hechas por otra persona. En caso que corresponda a "copia" a otros estudiantes, la sanción anterior se aplicará a todos los involucrados. En todos los casos, se informará a la Dirección de Pregrado de la Escuela de Ingeniería para que tome sanciones adicionales si lo estima conveniente.

También se entiende por copia extraer contenido sin modificarlo sustancialmente desde fuentes digitales como Wikipedia o mediante el uso de asistentes inteligentes como ChatGPT o Copilot. Se entiende que una modificación sustancial involucra el análisis crítico de la información extraída y en consecuencia todas las modificaciones y mejoras que de este análisis se desprendan. Cualquiera sea el caso, el uso de fuentes bibliográficas, digitales o asistentes debe declararse de forma explícita, y debe indicarse cómo el/la estudiante mejoró la información extraída para cumplir con los objetivos de la actividad evaluativa.

Obviamente, está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, **siempre y cuando se incluya la referencia correspondiente**.

Lo anterior se entiende como complemento al Reglamento del Estudiante de la Pontificia Universidad Católica de Chile (`https://registrosacademicos.uc.cl/reglamentos/estudiantiles/`). Por ello, es posible pedir a la Universidad la aplicación de sanciones adicionales especificadas en dicho reglamento.

Compromiso del Código de Honor

Este curso suscribe el Código de Honor establecido por la Universidad, el que es vinculante. Todo trabajo evaluado en este curso debe ser propio. En caso que exista colaboración permitida con otros/as estudiantes, el trabajo deberá referenciar y atribuir correctamente dicha contribución a quien corresponda. Como estudiante es un debe conocer el Código de Honor (`https://www.uc.cl/codigo-de-honor/`).