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

IIC2115 – Programación como Herramienta para la Ingeniería (II/2020)

Laboratorio 4 - Análisis exploratorio de datos

Objetivos

- Aplicar los contenidos de análisis exploratorio de datos para procesar información y tomar decisiones en base a esto.

Entrega

- **Lenguaje a utilizar:** Python 3.6
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **L04**.
- **Entrega: Domingo 8 de noviembre a las 23:59 hrs.**
- **Formato de entrega:**
  - Archivo python notebook (**L04.ipynb**) con su código. Utilice múltiples celdas de texto y código para facilitar la revisión de su laboratorio.
  - Archivo python (**L04.py**) con el mismo código entregado en el notebook.
  - Todos los archivos deben estar ubicados en la carpeta **L04**. No se debe subir ningún otro archivo a la carpeta. Los archivos **.ipynb** y **.py** deben contener la misma solución. No incluya las bases de datos en esta entrega.
- **Descuentos:** El descuento por atraso se realizará de acuerdo a lo definido en el programa del curso. Además de esto, tareas que no cumplan el formato de entrega tendrán un descuento de 0.5 ptos.
- **Laboratorios con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**

- Si su laboratorio es entregado fuera de plazo, tiene hasta el **Lunes 9 de noviembre a las 11:59 am** para responder el formulario de **entregas fuera de plazo** disponible en el Syllabus.
- Las discusiones en las *issues* del Syllabus en GitHub son parte de este enunciado.
- El uso de librerías externas que sean estructurales en la solución del los problemas no podrán ser utilizadas. Solo se podrán utilizar las que han sido aprobadas en las *issues* de GitHub.
- **Todos los comandos e instrucciones importantes de cada solución deben ser implementados usando funciones de *pandas*, *matplotlib*, *numpy* y/o *sklearn*. Esto significa que no está permitido iterar sobre los `DataFrame`, calculando con funciones básicas de Python lo requerido.**

Introducción

En este laboratorio utilizará el *set* de Datos Financieros disponible en la carpeta de este Laboratorio en el Syllabus. Este conjunto de datos consiste en más de 200 indicadores financieros para distintas acciones del mercado de acciones de los Estados Unidos. La información se encuentra disponible para cinco años separadas en cinco archivos de texto (`.csv`) diferentes (todos con las mismas columnas):

- 2014_Financial_Data.csv
- 2015_Financial_Data.csv
- 2016_Financial_Data.csv
- 2017_Financial_Data.csv
- 2018_Financial_Data.csv

Existen 2 tareas asociadas al *set* de datos, la primera consiste en predecir si las acciones son convenientes de comprar (columna `Class`, donde `Class` es 1 si es conveniente y 0 si no), mientras que la segunda consiste en predecir la variación en el precio de cada acción (columna `20XX PRICE VAR [%]`, donde `XX` es el año de la tabla). De los cinco archivos que conforman el *set* de datos, debe utilizar los correspondientes a los años entre 2014 y 2017 para entrenar, y el archivo correspondiente a 2018 para evaluar el rendimiento. El resto de columnas corresponden a descripciones, indicadores económicos y otros resultados de las empresas, que son fácilmente deducibles con una búsqueda por internet.

En base a los datos recién descritos y utilizando las librerías presentadas en clases, deberá cumplir una nueva serie de misiones relacionadas con el análisis exploratorio de datos.

Misiones

Para completar las Misiones, usted deberá responderlas con el objetivo de finalmente contar con los modelos de predicción de precio y conveniencia:

M1. **Conociendo e importando los datos:** Su primera misión será importar los datos provenientes de los cinco archivos de texto en `DataFrames` de la librería `Pandas`. Con los `DataFrames` ya creados, visualice los tipos de datos y consulte algunos estadísticos generales con los métodos revisados en clases. Para completar esta misión deberá presentar previsualizaciones (a modo tabla) de los datos. Tenga en cuenta que dado que las tablas poseen muchas columnas, es su objetivo previsualizar un subconjunto de solo 10 de estas. Agregue un párrafo de comentarios generales en una celda de texto sobre los estadísticos observados. **(1 pto.)**

M2. **Limpieza y depuración:** Analice la existencia de valores por defecto, nulos, incompletos, *outliers* o posibles inconsistencias. Utilice técnicas de datos para limpiar o imputar este tipo de anomalías. Para completar esta misión deberá mostrar cinco resoluciones de conflicto (valores por defecto, nulos, incompletos, *outliers* o posibles inconsistencias) variadas. Por ejemplo, aplicar algún modelo o regresión en la corrección, agrupar para corregir nulos, entre otros. Cabe destacar que si usted considera resolver más de cinco columnas, es libre de hacerlo en base a las necesidades de sus futuros modelos. **(1 pto.)**

M3. **Visualización:** Realice un análisis de datos por visualización. Es decir, construya gráficas o tablas interesantes que permitan comprender la naturaleza y relación de los datos presentes. Un ejemplo, es el presentado en la materia del curso con los datos crediticios. En esta misión puede agrupar o relacionar relacionar los datos. Para cumplir esta misión deberá presentar al menos 5 visualizaciones interesantes y diferentes. Si su trabajo es original y tiene consistencia con el análisis, puede recibir una bonificación extra. **(1 pto.)**

M4. **Entrenamiento de modelos parte 1:** Construya modelos de clasificación y/o regresión para identificar las acciones convenientes de comprar (basado en la columna `Class`). Encuentre el mejor modelo usando correctamente el proceso de validación. En una celda de texto comente el rendimiento encontrado en cada modelo. **(1 pto.)**

M5. **Selección de mejores *features*:** Con el objetivo de predecir la variación del precio de las acciones (columna `20XX PRICE VAR [%]`), identifique las mejores 10 *features* a considerar en la estimación de un modelo. Esta misión es completamente interactiva con las misiones 2 y 6, por lo que podrían ir

modificándose a medida que avanza en este laboratorio. Justifique debidamente qué criterios utiliza para decir que una *feature* es buena o no para realizar la regresión. **(1 pto.)**

M6. **Entrenamiento de modelos parte 2:** Estime y utilice modelos para predecir valores de acciones a futuro, permitiendo predicciones año a año hasta un horizonte predefinido. En base a esto, encuentre las cinco mejores acciones a comprar y las cinco peores para el año 2030. **(1 pto.)**

Corrección

Es importante que deje ejecutado todo su trabajo antes de subirlo, de lo contrario se le aplicará un descuento al puntaje total. Para la corrección de este laboratorio, se revisarán los procedimientos desarrollados para responder las diferentes misiones propuestas y la estructura de como utiliza los módulos *pandas*, *matplotlib*, *numpy* y/o *sklearn* en ellos. 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

"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.