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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2022)

Laboratorio 2

Aspectos generales

- **Objetivo:** evaluar individualmente el aprendizaje sobre análisis y visualización de datos en Python, a través de la construcción de una serie de tareas asociadas las estadísticas de las ligas de fútbol europeo.
- **Lugar de entrega:** domingo 8 de mayo a las 23:59 hrs. en repositorio privado.
- **Formato de entrega: UNICAMENTE** el archivo Python Notebook (**L2.ipynb**) con la solución del laboratorio. El archivo debe estar ubicado en la carpeta **L2**. Es requerimiento de formato el utilizar múltiples celdas de texto y código para la construcción de la solución. Laboratorios 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. Si su laboratorio es entregado fuera de plazo, tiene hasta el **lunes 9 de mayo a las 11:59 AM** para responder el formulario de **entregas fuera de plazo** disponible en el Syllabus.
- **Issues:** Las discusiones en las *issues* del Syllabus que sean relevantes para el desarrollo del laboratorio, serán destacadas y se considerarán como parte de este enunciado. Así mismo, el uso de librerías externas que solucionen aspectos fundamental del laboratorio no podrán ser utilizadas. Solo se podrán utilizar las que han sido aprobadas en las *issues*, previa consulta de los estudiantes.
- **Laboratorios con errores de sintaxis y/o que generen excepciones en todas las ejecuciones serán calificados con nota 1.0.**

Introducción

En este laboratorio utilizará un conjunto de datos que recopila información de 12.290 partidos de diversas ligas de fútbol europeo entre los años 2014 y 2020. Estos datos consideran los siguientes indicadores por cada equipo participante en cada partido:

- **league**: liga europea a la que corresponde el partido.
- **season**: año en que comenzó la temporada donde se realizó el partido.
- **data**: fecha en que se jugó el partido.
- **team**: equipo para el cual se indican los datos del partido.
- **h_a**: localía (h) o visita (a) del equipo analizado.
- **result**: resultado del partido para el equipo analizado: triunfo (w), derrota (l), empate (d).
- **pts**: puntos obtenidos en el partido por el equipo analizado.
- **goals_scored**: goles anotados por el equipo analizado.
- **goals_missed**: goles recibidos por el equipo analizado.
- **deep_passes**: pases exitosos realizados por el equipo analizado en los últimos 20 metros de la cancha (pases profundos).
- **deep_passes_allowed**: pases profundos exitosos realizados por el contrincante.
- **ppda**: índice de presión defensiva, calculado a partir de los pases realizados por el equipo contrincante por cada acción defensiva realizada por el equipo analizado (mientras más bajo, más presión).
- **oppda**: índice de presión defensiva del equipo contrincante.

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

Misiones

Todas las misiones son independientes entre sí y tienen el mismo puntaje, por lo que puede realizarlas en el orden que prefiera:

1. **Carga y exploración de los datos**: cargue los datos contenidos en el archivo `data.csv` en un `DataFrame`, obtenga los tipos de datos de cada columna y consulte algunos estadísticos generales con los métodos revisados en clases. A continuación, presente visualizaciones relevantes para las columnas, agregando un párrafo de comentarios para cada una, analizando los estadísticos observados y caracterizando la distribución.

2. **Agregación**: utilizando los datos, construya un nuevo `DataFrame` con información agregada, que resuma la tabla de posiciones de cada liga para cada año. El `DataFrame` debe presentar al menos la siguiente información por cada registro:

   - **league**: liga europea a la que corresponde el equipo.
   - **season**: año en que comenzó la temporada.
   - **position**: posición final en la liga.
   - **team**: equipo analizado.
   - **matches**: número de partidos disputados durante la temporada.
   - **pts**: puntos obtenidos durante la temporada.
   - **wins**: número de partidos ganados durante la temporada.
   - **draws**: número de partidos empatados durante la temporada.
   - **defeats**: número de partidos perdidos durante la temporada.
   - **goals_scored**: goles anotados durante la temporada.
   - **goals_missed**: goles recibidos durante la temporada.
   - **deep_passes_avg**: promedio de pases profundos logrados por partido durante la temporada.
   - **deep_passes_allowed_avg**: promedio de pases profundos permitidos por partido durante la temporada.
   - **ppda_avg**: promedio del índice de presión defensiva durante la temporada.
   - **oppda_avg**: promedio del índice de presión defensiva de los equipos contrincantes durante la temporada.

El `DataFrame` debe estar ordenado, respetando el siguiente nivel de importancia: liga, año (ascendente), posición (ascendente). Como referencia, considere la siguiente tabla, que presenta los primeros cinco registros de un `DataFrame` que cumple lo solicitado:

league	season	position	team	matches	pts	wins	draws	defeats	goals_scored	goals_missed	deep_passes_avg	deep_passes_allowed_avg	ppda_avg	oppda_avg
La_liga	2014	1	Barcelona	38	94	30	4	4	110	21	12.868421	3.000000	5.683535	16.367593
La_liga	2014	2	Real Madrid	38	92	30	2	6	118	38	9.236842	4.026316	10.209085	12.929510
La_liga	2014	3	Atletico Madrid	38	78	23	9	6	67	29	5.184211	3.236842	8.982028	9.237091
La_liga	2014	4	Valencia	38	77	22	11	5	70	32	5.342105	4.526316	8.709827	7.870225
La_liga	2014	5	Sevilla	38	76	23	7	8	71	45	8.026316	4.421053	8.276148	9.477805

3. **Visualización**: basándose en el índice `ppda` y en un análisis visual, proponga una metodología que permita identificar los cambios de director técnico en los equipos. Esta metodología no tiene que ser infalible ni perfecta, pero debe tener una justificación razonable desde el punto de vista del análisis estadístico realizado. Muestre al menos 3 casos de aplicación de la metodología.

4. **Predicción de resultados**: en base a los datos disponibles, construya un modelo que permita predecir las probabilidades de los resultados de un partido. En particular, dados ambos equipos y la especificación de localía, se espera que el modelo calcule la probabilidad de que gane el local, la visita o haya un empate. Independiente de la metodología y técnicas utilizadas para construir el modelo, entregue métricas de validación del rendimiento de este.

5. **Caracterización de las ligas**: en base a los datos disponibles, realice un análisis que permita caracterizar y diferenciar a las distintas ligas. No es necesario que este análisis genere una métrica única de caracterización.

6. **Fútbol ofensivo o defensivo**: basándose en los datos, intente responder a una de las preguntas más antigua del fútbol, ¿qué equipos tienen más éxito, los ofensivos o los defensivos?

* **(Bonus)**: simule una temporada de una liga de su elección, indicando los resultados (no los marcadores) de cada partido y la tabla de posiciones final.

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 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

Los alumnos de la Escuela de Ingeniería deben mantener un comportamiento acorde al Código de Honor de la Universidad:

"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.