Proyecto de Estadística Aplicada III (2020-II)

Plan de trabajo para proyectos.

Introducción

Como parte de llevar a la práctica los métodos y técnicas desarrolladas en clase, se debe realizar un proyecto en el que se apliquen una o más de las técnicas aprendidas para resolver un problema y generar un reporte de análisis, junto con una presentación que puede o no presentarse ante el grupo. Con la finalidad de hacer el proyecto interesante, se podrá utilizar un conjunto de datos propuesto por el equipo, o bien, utiizar alguno de los conjuntos de datos disponibles en algún repositorio como:

- Tidytuesday,
- Kaggle,
- CRedit Risk Analytics
- UCI Machine Learning Repository.
- Yahoo Sandbox datasets.
- Jerry Smith Dataset Collection
- Sports Statistics
- SourceForge.net Research Data

Los anteriores son sólo unos posibles entre otros posibles muchos (citar en caso de usar otro repositorio)

El procedimiento para realizar el proyecto es el siguiente:

1. Definir equipos de trabajo. Aunque los proyectos pueden ser individuales, se considera que el análisis será mejor discutido en equipos de a lo más 4 personas. No hay excepciones a esta regla.

2. Se deberá realizar una propuesta de proyecto que el instructor tiene que aprobar, en donde se defina la motivación del tema, el alcance del proyecto, los resultados que se espera obtener, las actividades que se realizarán y las metodologías que se planea usar. La propuesta podrá tener de una a dos páginas, *y será parte del trabajo final.*

3. El proyecto deberá quedar definido a más tardar el día viernes 6 de noviembre de 2020 . L@s alumn@s que no tengan un proyecto definido para esa fecha tendrán que hacer el que les defina el instructor.

4. El proyecto será revisado y devuelto con comentarios a más tardar el 9 de noviembre por mi parte.

5. El proyecto se entregará el último día de clases. Lo que se entregará es el documento del proyecto junto con una presentación que es la que se presentaría a una audiencia sobre las ideas principales del proyecto (pensando en una presentación de 20 minutos). **Aunque ésta es posible que no se lleve a cabo, es parte del entregable**.

Proyectos

Dentro de los conjuntos de datos disponibles en Kaggle, *se sugiere* considerar los siguientes. Algunos no cuentan con un problema, son simplemente datos disponibles:

- How do you measure justice?: How do you measure justice? And how do you solve the problem of racism in policing? We look for factors that drive racial disparities in policing by analyzing census and police department deployment data. The ultimate goal is to inform police agencies where they can make improvements by identifying deployment areas where racial disparities exist and are not explainable by crime rates

and poverty levels. Our biggest challenge is automating the combination of police data, census-level data, and other socioeconomic factors. Shapefiles are unusual and messy -- which makes it difficult to, for instance, generate maps of police behavior with precinct boundary layers mixed with census layers. Police incident data are also very difficult to normalize and standardize across departments since there are no federal standards for data collection.

- Focused customer Churn: Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs.
- SF Police Calls for service and Incidents: This is a dataset hosted by the city of San Francisco. The organization has an open data platform found here and they update their information according the amount of data that is brought in. Explore San Francisco's Data using Kaggle and all of the data sources available through the San Francisco organization page!
- Los Angeles Metro Bike Share Trip data: This is a dataset hosted by the city of Los Angeles. The organization has an open data platform found here and they update their information according the amount of data that is brought in. Explore Los Angeles's Data using Kaggle and all of the data sources available through the city of Los Angeles organization page! Update Frequency: This dataset is updated daily.
- Precios de Aguacate en diferentes mercados: It is a well known fact that Millenials LOVE Avocado Toast. It's also a well known fact that all Millenials live in their parents basements. Clearly, they aren't buying home because they are buying too much Avocado Toast! But maybe there's hope... if a Millenial could find a city with cheap avocados, they could live out the Millenial American Dream.
- LA Restaurant & Market Health data: This dataset contains Environmental Health Inspection Results for Restaurants and Markets in the City of Los Angeles. Los Angeles County Environmental Health is responsible for inspections and enforcement activities for all unincorporated areas and 85 of the 88 cities in the County. This dataset is filtered from County data to include only facilities in the City of Los Angeles. The full dataset is available at here.
- Genetic Variant Classifications: Predict whether a variant will have conflicting clinical classifications. ClinVar is a public resource containing annotations about human genetic variants. These variants are (usually manually) classified by clinical laboratories on a categorical spectrum ranging from benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (from laboratory to laboratory) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient.
- Wine Reviews: How create a predictive model to identify wines through blind tasting like a master sommelier would.

La forma en que se evaluará incluirá, entre otras cosas, los siguientes elementos: - La definición adecuada del problema que se pretende resolver, con sus respectivas hipótesis - La adecuación de la metodología usada al problema planteado. - Las herramientas utilizadas (cualesquiera que estas sean, no se busca enfatizar el uso en R). - La porcentaje de datos utilizados con respecto a la disponibilidad de datos en la base original.

Proyecto basado en un artículo

Adicional a los temas sugeridos, se pondrán en Piazza una sección de papers para proyectos que incluyen diversos temas. La idea es reproducir el paper, aplicarlo a un conjunto específico de datos, elaborar un laboratorio de aplicación, ampliar las ideas del paper, elaborar un tutorial, adaptarlo a una situación particular, elaborar un programa, etc.

El artículo provee de ideas y de posibles ambientes en donde se puede repasar o ampliar los temas de estudio del curso. Se debe seguir la misma idea que en los proyectos basados en datos que se menciona arriba.