49-733 / 90-835 Designing Smart and Healthy Systems (Fall 2021)
HW #1: Exploratory Data Analysis and Data Preparation
<u>Due November 2, 2021</u>
David Steier

The focus of this assignment is on exploratory data analysis and data preparation. You will practice skills of profiling, transforming, and cleansing the data prior to use in predictive modeling. Each team will get a set of input files to profile, a set of care events to create from the input data for use as target variables, and a subset of the predictor variables to create from the input data sets. This is a group assignment for your team. Please submit one assignment for each team (and be sure to include your team name and list the team members on your submission). Files turned in may be a combination of .pdf files (for hypotheses) and Jupyter notebooks.

Data

The data for this assignment (and the remainder of the course) is stored in the S3 data bucket `cmu-dshs-diabetes-data`. Check out the notebook file `Patient EDA.ipynb` for an example of how to access the data in the first file.

The S3 bucket contains eight files

- `Balance – Tablas (Parte 1)`: patient care data, with main tabs including agenda (appointments), hc (patient history), hc_hospitalizaciones (hospitalizations), historialcitascanceladas (cancellations), historialcitasagendadas (history of scheduled appointments), pacientes (patients), IdPaquete (package ID), periodosrecetas (period codes), pr_razones (reason codes), prospectos (prospects)
- `Balance – Tablas (Parte 2)` : patient care data, with tabs Prospectosseguimiento (prospect followups), psicologiasesion1 (psychology session 1)
- `Balance – Tablas (Parte 3`): patient care data with tabs visitas (visits), valoracionpodologicaseguimiento (follow-up assessments), valoracionesexpress (express assessments/ratings)
- `I Datos CMU (BAL)`: main tabs bitacoravisitas (visiit records), consultas (consultations), controlvital (vital signs), laboratoriosotros (other labs), notasevolucionpsic (psychological progress notes?),
- `II Datos CMU (BAL)`: with: main tabs nutriciondietaspacientes (patient nutrition diets), nutricionevaluacionnutricional (nutritional assessment), nutricionnotasseguimiento (nutrition follow-up notes), nutricionplanesnutricionales (nutrition plans), nutricionvaloracioninicial (initial nutrition assessments), testhamilton (Hamilton test), retinografias (retinographies)
- `III Datos CMU (BAL)`: main tab encuestas_servicios (service survey?)
- `IV Datos CMU (BAL)`: main tabs cm_tickets (membership), evoluciones (progress?),
- `DataDictionary`: translations of fields grouped under controlvital, evoluciones, consultas, bitacoravisitas, hc, retinografias, and laboratorios

Assignment tasks

Each task is worth 5 points for a total of 20 points for the assignment.

1. Each team should create ten hypotheses to test about relationships between potential predictors and diabetes patient engagement. An example might be that patients who live close to a clinic are more likely to be highly engaged in their treatment. You may use the readings, information gathered in class, searching the Web or talking to subject matter experts to generate your hypotheses.
2. Each group should profile the data in the files using Python as follows:
    - **Group 1**: `Balance – Tablas (Parte 1)`
    - **Group 2**: `Balance – Tablas (Parte 2)` and `Balance – Tablas (Parte 3)`
    - **Group 3**: `I Datos CMU (BAL)` and `II Datos CMU (BAL)`
    - **Group 4** `III Datos CMU(BAL`) and `IV Datos CMU (BAL)`

Profiling should include at a minimum the number of records in a file, the fields in a record, and for each field, the format of the data, units, the range of values, and summary statistics for numeric data. You should create some data visualizations to understand the range of values and spot outliers. If you are not familiar with data preparation steps, a good reference is Chapter 3, "Data Preprocessing" in J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques*, Third Edition, Morgan Kaufmann, 2012 (see slides in https://hanj.cs.illinois.edu/bk3/bk3_slides/03Preprocessing.ppt)

3. Each group should derive a subset of the care events (engagement outcomes) from the data files as follows
   - **Group 1**
     - Skips appointment, from the input file Balance – Tablas (Parte 1)
     - Cancels appointment, from the input file Balance – Tablas (Parte 1)
     - Reschedules appointment, from the input file Balance – Tablas (Parte 1)
   - **Group 2**
     - Attends Single-Service Appointment from the input file Balance – Tablas (Parte 1)
     - Attends All-Service Appointment, from the input file Balance – Tablas (Parte 1)
   - **Group 3**
     - Complains from the input file III Datos CMU (BAL)
     - Buys from Clinic, from the input file IV Datos CMU (BAL)
   - **Group 4**
     - Attends Single-Service Appointment and Buys from Clinic from the input files Balance – Tablas (Parte 1) and IV Datos CMU (BAL)
     - Attends All-Service Appointment and Buys from Clinic, from the input files Balance – Tablas (Parte 1) and IV Datos CMU (BAL)
     - Buys or Renews Membership, from the input files IV Datos CMU (BAL)
     - Cancels Membership, from the input files IV Datos CMU (BAL)
4. Each group will derive a subset of the engagement predictor variables in the following categories
   - **Group 1**
     - Demographics, from the input files Balance – Tablas (Parte 1). For example marital status could be derived from EstadoCivil
     - Geography, from the input files Balance – Tablas (Parte 1). For example, distance from clinic is dependent on city and state which are in Ciudad and Estado respectively
   - **Group 2**
     - Social environment, from the input files Balance – Tablas (Parte 1). Social support could be inferred from the psychological notes
     - Mental health background, Balance – Tablas (Parte 1). The presence of mental illness can be found in the psychological notes
   - **Group 3**
     - Medical history, from the input files Balance – Tablas (Parte 1). Various medical conditions are listed in the hc and hc_hospitalizaciones tabs
   - **Group 4**
     - Lab findings, from the input file Balance – Tablas (Parte 3)
     - Physical exam findings: from the input Balance – Tablas (Parte 3)
     - Lifestyle and nutrition, from the input file II Datos CMU (BAL)
     - Eye care exam results, from the input file II Datos CMU (BAL). For example the diagnosis of diabetic retinopathy is available on the retinographias tab

You should be on the lookout for risks around data that might impede your ability to create a good model. These include data that are:

- Incomplete
  - Records missing

    - Fields missing or not populated
- Inconsistent
    - Same field means different things in different sources
    - Attributes in one source are relations in another source
    - Different data formats are used in different sources
    - Different values represent the same object
- Unlinked – keys do not match across tables
- Inaccurate
    - Data is unintentionally or intentionally corrupted (or wrong to begin with)
    - Data was originally accurate, but now out of date
    - Data has insufficient precision/granularity