PROJECT 2:
SPACESHIP TITANIC

IE 528

Data Analytics & Mining

TITANIC - ML FROM DISASTER

- Predict whether a passenger was transported to an alternate dimension during the Spaceship Titanic's collision with the spacetime anomaly
  - Source: https://www.kaggle.com/competitions/spaceship-titanic
  - ~13000 instances (~8700 train + ~4300 test) with 13 attributes (some have missing values) and one binary target (*Transported*)
    - Balanced (n(True) = 4378, n(False) = 4315, n(Total) = 8693 in train set)
  - (Preliminary) Data Pre-processing
    - *Cabin* needs to be split into three vars (*deck*/*num*/*side*); *num* may be dropped.
    - *PassengerId* and *Name* may be dropped; however, *Name* may be used to extract family name which can help to identify "*family size*."
    - *RoomService*, *FoodCourt*, *ShoppingMall*, *Spa*, *VRDeck* may be integrated into a single "*Billing*" feature, which may be related to *VIP*.
- Follow the same steps as the previous project (see the next slide)
  - Use the train set only; CV error is the KPM (~80% or less)
- Submit a MATLAB live script (.mlx)
  - Use the mlx provided for the previous project as the template

TASKS

1. Data preparation: Read *Overview* & *Data* and download the dataset.
   - Provide a brief overview about the dataset and description about included attributes
2. Exploratory Data Analysis (EDA) and Data Preprocessing
   - Analyze the raw data using data summary and visualization (e.g., histograms, scatter plots, correlation plots, heatmaps, etc.)
   - Identify relationships between attributes and target, handle missing values, transform data or apply encoding as needed, create new features or drop irrelevant ones, etc.
3. Training classifiers
   - Use *any* techniques you've learned so far: (DT, NB, kNN,) ANN, SVM, Ensemble
   - Train and evaluate the models by CV (please set rng('default') for reproducibility!)
   - Tune hyper-parameters to obtain the best model from each type
     - For example, # hidden neurons in ANN, kernel scale in SVM, # learning cycles in ensemble, etc.
4. Evaluation
   - Since we don't know truth values for the test set, let's use ~~resub &~~ (10-fold) CV error.
     - 20.5%/19.3%/22.7% resub and 22.1%/26.1%/28.2% CV error by vanilla DT/ANN/SVM (w/o *PassengerId, Name, Cabin*)
   - Also obtain a confusion matrix for each model and calculate F1.

EVALUATION

- Requirements
  - 3 members in each group; only one member must submit the work.
  - Use only MATLAB. DO NOT use other tools or languages.
  - Submit the following file:
    - MATLAB live script (.mlx) which contains the code with description for each section/task and the result
    - The code should read train.csv and generate everything that is needed in the workspace. Therefore, MAT is not needed.
  - Every member can submit a peer evaluation form (optional).

- Evaluation criteria:
  - Data processing (20%), model quality (20%), performance (20%), report (live script) clarity (30%), peer eval (10%)

DATA PRE-PROCESSING FOR ANN

- Option 1: train(patternnet)
  - Explicit one-hot encoding is needed
  - T = readtable('train.csv');
  - X = T(:,2:12); % excluding PassengerId & Name; Name might be useful
  - Y = T(:,14);
  - XH = **categorical**(X.HomePlanet); OXH = **onehotencode**(table(XH));

    % Repeat for one-hot encoding of all categorical variables
  - NX = [OXH, OXC, ..., table(X.Age, X.RoomService, ...)];

    % Each element must be a table
  - NX = table2array(NX); % Input data for train() must be in an array
  - NY = **dummyvar**(Y.Y); % use either onehotencode() or dummyvar()
- Option 2: fitcnet()
  - Similar to other fitcxxx(); tables with numeric & categorical can be used
    - Automatically find if a var in table is categorical (check Mdl's CategoricalPredictors)
    - No explicit one-hot encoding is needed
  - Refer to https://www.mathworks.com/help/stats/fitcnet.html