Лабораторная работа: Анализ данных о пассажирах Титаника

В рамках лабораторной работы необходимо провести ручной анализ данных о пассажирах затонувшего корабля Титаник. Будут оцениваться четкость ответов на вопросы, аккуратность отчета и кода.

Задание

1) Загрузите набор данных: файл train.csv на странице учебного соревнования: Titanic: Machine Learning from Disaster.

Там же вы найдете описание формата файла. Для загрузки файлов необходимо зарегистрироваться в Kaggle, если вы еще не имеете аккаунта. Обратите внимание — train.csv содержит не всех пассажиров, меньшая их часть была отложена на контроль test.csv, но можно считать, что разбиение было случайным (попадание пассажира в ту или иную группы не зависит от пассажира и его характеристик).

2) Что данные могут рассказать о задаче? Попробуйте содержательно ответить на следующие вопросы, подкрепляя их свидетельствами в виде графиков, гистограмм, приводя статистики распределений (средние, медианы, стандартные отклонения):
   - Как много пассажиров ехало первым классом?
   - Какой возраст имели пассажиры?
   - Какое распределение детей по классам?
   - Коррелируют ли число братьев/сестер с числом родителей/детей?
   - Какие самые популярные имена на корабле?
   - Как варьируется цена билета на Титаник?
   - Какие титулы имели пассажиры (примеры титула — Mrs., Mr.), едущие различными классами?
   - Есть ли зависимость между классом и номером билета?
   - Какой части пассажиров удалось выжить?
3) Попробуйте вручную найти закономерности, описывающие выживших пассажиров:
   - Верно ли, что женщины выживали чаще мужчин?
   - Верно ли, что чаще выживали пассажиры с более дорогими билетами?
   - Найдите закономерности, точно описывающие группу выживших пассажиров (все пассажиры, попадающие под правило — выжили). Что можно сказать про сложность и интерпретацию этих закономерностей? Есть ли среди них логичные? Есть ли примеры ложных закономерностей?

Отчет

Результат лабораторной работы — **отчет**. Его нужно выполнить в формате ноутбуков IPython (ipynb-файл).

Сам код в отчете не столь интересен. Чем меньше кода, тем лучше всем: мне — меньше проверять, вам — проще найти ошибку или дополнить эксперимент.

Постарайтесь сделать ваш отчет интересным рассказом, последовательно отвечающим на вопросы из задания. Не забывайте подписывать оси на графиках!

Инструменты для выполнения задания

IPython Notebook
- Галерея интересных ноутбуков

Библиотеки Python
- NumPy
  - руководство для пользователей Matlab
- Pandas
  - пример работы с данными при помощи pandas
- Matplotlib
  - pyplot — эмуляция функционала графопостроений в Matlab
  - галерея примеров
- SciPy