Домашнее задание 1
Машинное обучение, БИБ21

1 Задание 1: Титаник

- Открыть в Pandas файл **titanic.csv** (см. вложения). Ответить на вопросы ниже, используя средства языка Python и необходимых библиотек;
- Какое количество мужчин и женщин ехало на корабле?
- Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров.
- Какую долю пассажиры первого класса составляли среди всех пассажиров? Постройте круговую диаграмму по всем классам пассажиров.
- Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров. В чем статистическая разница между показателями медианы и среднего? Выведите любой перцентиль возраста пассажиров. Что показывает перцентиль в статистике?
- Коррелируют ли число братьев/сестер/супругов с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch. Что показывает корреляция?
- Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name). Это задание — типичный пример того, с чем сталкивается специалист по анализу данных. Данные очень разнородные и шумные, но из них требуется извлечь необходимую информацию. Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские. Попробуйте написать скрипт для извлечения самого популярного женского имени как можно оптимальнее (меньше циклов, больше использований средств Pandas...);

2 Задание 2: Аппроксимация

Дана сложная математическая функция:

$$f(x) = \sin\left(\frac{x}{5}\right) * \exp\left(\frac{x}{10}\right) + 5\exp\left(\frac{-x}{2}\right),\ x \in [1, 15]$$

Необходимо найти такую функцию $g(x)$, что $f(x) \approx g(x)$, но при этом $g(x)$ имеет более простую форму. Очевидно, что можно выбрать многочлен в качестве функционала g. Итак, пусть $g(x)$ является многочленом n степени ($n > 0$).

Иначе говоря, задача состоит в поиске функции вида

$$f(x) \approx g(x) = a_0 + a_1 x + ... + a_n x^n.$$

Так как любой полином однозначно определяется по любым $n + 1$ различным точкам, через которые он проходит, то для нахождения набора коэффициентов $a_i{}_{i=0}^{n}$ необходимо решить следующую СЛАУ:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + ... + a_n x_0^n = f(x_0) \\ a_0 + a_1 x_1 + a_2 x_1^2 + ... + a_n x_1^n = f(x_1) \\ ... \\ a_0 + a_1 x_n + a_2 x_n^2 + ... + a_n x_n^n = f(x_n) \end{cases}$$

- Сформируйте систему линейных уравнений (то есть задайте матрицу коэффициентов A и свободный вектор b) для многочлена первой степени, который должен совпадать с функцией f в точках 1 и 15. Решите данную систему с помощью функции **scipy.linalg.solve**. Нарисуйте функцию f и полученный многочлен. Хорошо ли он приближает исходную функцию?

- Повторите те же шаги для многочлена второй степени, который совпадает с функцией f в точках 1, 8 и 15. Улучшилось ли качество аппроксимации?

- Повторите те же шаги для многочлена третьей степени, который совпадает с функцией f в точках 1, 4, 10 и 15. Хорошо ли он аппроксимирует функцию? Коэффициенты данного многочлена (четыре числа: w_0, w_1, w_2, w_3) являются ответом на задачу.

- Приведите примеры задач, для которых необходима аппроксимация. Какие метрики существуют для определения эффективности аппроксимации? Приведите пример.