# **Построение объясняющих моделей ИИ (XAI)**

**Цель** работы – получить навыки создания систем объяснений для интерпретации моделей машинного обучения, а также визуализации объяснений работы моделей для сущностей из наборов данных.

**Задачи** – Построить следующие системы объяснений для двух моделей (одного классификатора и одного регрессора), полученных при выполнении работ ранее: графики частичной зависимости (PDP – partial dependence plot), важность перестановок (permutation importance) значения Шепли (Shapley values).

Выполнить загрузку моделей, полученных ранее (классификатор и регрессор). Загрузить выборки данных, использовавшихся для обучения выбранных моделей.

Выполнить следующие задачи для каждой модели:

1. PDP: Построить графики индивидуального условного ожидания (ICE) для всех входных признаков для набора 30-50 записей из оригинальной выборки и средний график частичной зависимости. Выбрать три признака, оказывающие наибольшее влияние на целевую переменную. Для каждой комбинации из двух важных признаков, входящих в Топ-3, построить двумерный график частичной зависимости, а для двух наиболее важных – трехмерный график частичной зависимости.

2. Важность перестановок: Построить диаграммы размаха по всем входным признакам исходя из двух метрик качества, соответствующие моделям.

3. Значения Шепли: Получить итоговый график важности признаков для модели (summary\_plot) – значения Шепли для 200-500 записей в виде beeswarm (SHAP Value (impact on model output)) и barplot (mean(|SHAP|) (average impact on model output magnitude)). Вывести два графика зависимости (dependence plot): для комбинации наиболее важного признака и второго по важности признака, для комбинации наиболее важного признака и третьего по важности признака. Построить график решения (decision plot) для трех записей из набора. Построить силовой график (force plot) для двух других записей из набора.

Cделать выводы о важности признаков, влияющих на целевую переменную, исходя из результатов всех инструментов.

**Описание отчета.**

1. Описание наборов данных.

2. Графики по заданию

4. Оценки важности признаков для каждого из инструментов

5. Программный код с комментариями.

6. Выводы