**Министерство науки и высшего образования Российской Федерации**
**Федеральное государственное бюджетное образовательное учреждение высшего образования**
**«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»**
**Факультет: Робототехника и комплексная автоматизация**
**Кафедра: Системы автоматизированного проектирования**

# Интерпретация аппроксимирующих моделей машинного обучения

Выполнил: Конов А.В.
Научный руководитель: д.ф-м.н., профессор, Карпенко А.П.

# Введение

Интерпретация аппроксимирующих моделей машинного обучения - это процесс объяснения того, как модель делает свои предсказания. Интерпретация включает анализ влияния признаков на предсказания модели, выявление зависимостей и предоставление понятных объяснений для решений, принимаемых моделью. Цель - обеспечить понимание и доверие к модели, а также помочь выявить причинно-следственные связи и использовать модель для принятия решений.

# Цели и задачи

Цель работы – разработка ПО для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения на основе существующих методов интерпретации.

Задачи:

- Изучить аппроксимирующие модели машинного обучения;
- Изучить методы интерпретации аппроксимирующих моделей машинного обучения;
- Предложить алгоритм для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения;
- Реализовать предложенный алгоритм и исследовать его эффективность.

# Постановка задачи машинного обучения

В общем виде задача машинного обучения выглядит следующим образом:

$$f^* = \arg\min_f L(f, D),$$

где $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ - обучающая выборка; $x_i$ - вектор признаков; $y_i$ - целевая переменная для соответствующего вектора признаков; $f$ - модель, аппроксимирующая зависимость между признаками и целевой переменной, $L$ - функция потерь, оценивающая расхождение между предсказанными и фактическими значениями целевой переменной на обучающей выборке; $f^*$ - оптимальная модель.

*Вектор признаков* - это числовое представление объекта, где каждая компонента вектора соответствует определенному признаку и описывает его характеристику или значение.
*Целевая переменная* - это значение, которое модель машинного обучения пытается предсказать или оценить на основе доступных признаков и обучающих данных.
*Обучающая выборка* - это набор примеров данных, которые используются для обучения модели машинного обучения, состоящий из входных признаков и соответствующих им целевых переменных или меток.

# Модели машинного обучения

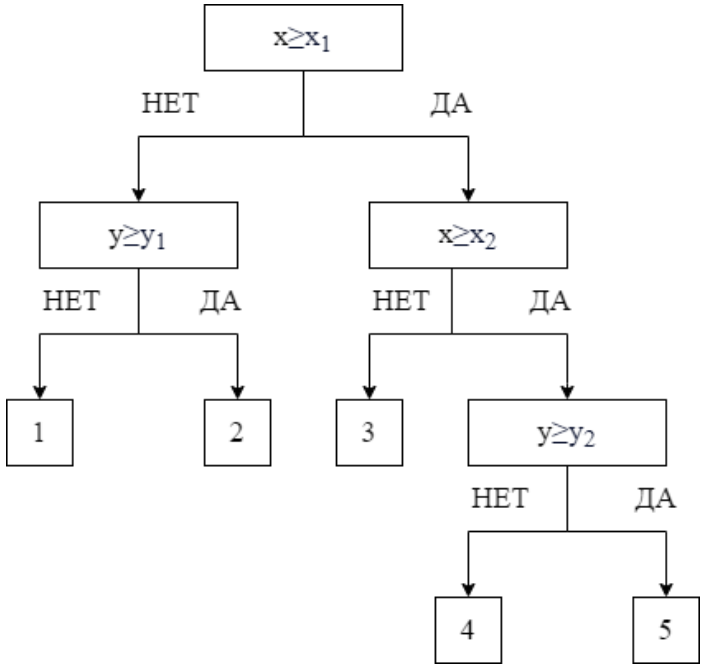
Классические:
- Линейная регрессия
- Логистическая регрессия
- Метод опорных векторов
- Наивный байесовский классификатор

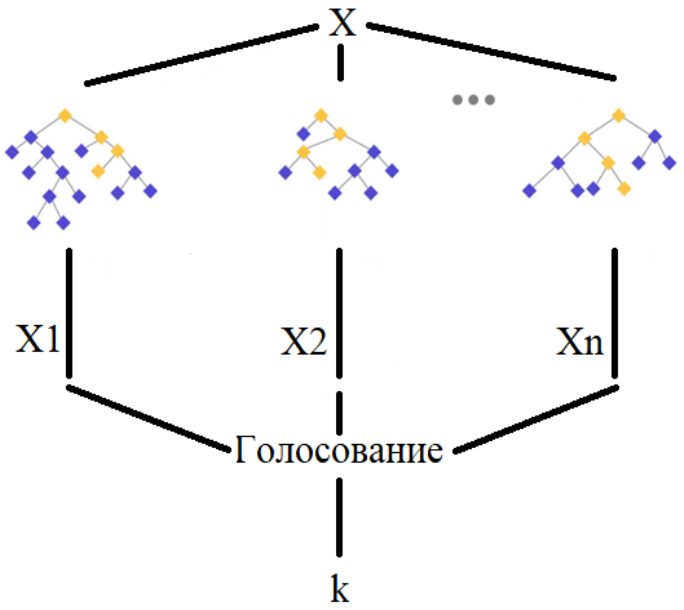
На основе решающих деревьев:
- Решающие деревья
- Случайный лес
- Бустинг

Нейронные сети:
- Прямого распространения
- Свёрточные
- Реккурентные

Модели на основе решающих деревьев обладают рядом преимуществ перед остальными моделями:
- хорошо интерпретируемы;
- могут обрабатывать категориальные и числовые значения;
- хорошо подходят как для работы с малым, так и для работы с большим количеством данных;
- устойчивы к выбросам;
- обладают высокой скоростью обучения.

# Модели машинного обучения. Решающее дерево и случайный лес




Решающее дерево представляет собой иерархическую структуру в виде дерева, в которой каждый узел представляет условие, а каждый листовой узел представляет класс или числовое значение для задач регрессии.

Случайный лес - ансамбль решающих деревьев, которые обучаются независимо и комбинируются путем голосования для получения более точных прогнозов, устойчивых к переобучению и способных обрабатывать разнообразные типы данных.

# Методы интерпретации моделей машинного обучения

Локальные:

- LIME
- ICE

Глобальные:

- SHAP
- PDP
- ALE

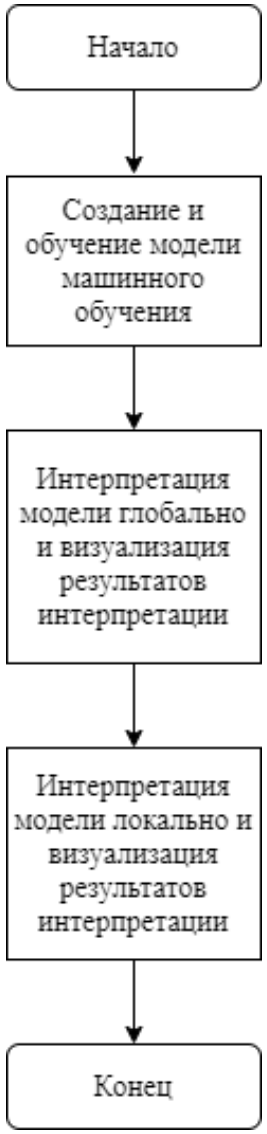
Локальные методы интерпретации относятся к методам, которые позволяют анализировать, какие признаки были наиболее важны для принятия решения моделью для конкретного объекта данных.

Глобальные методы интерпретации относятся к методам, которые позволяют анализировать важность признаков в модели в целом.

# Методы интерпретации моделей машинного обучения. Описание

Суть метода LIME заключается в создании локальной модели, которая объясняет прогноз модели на отдельном объекте.
Метод SHAP использует теорию кооперативных игр для объяснения важности каждого признака в модели, учитывая его взаимодействие с другими признаками.
В основе метода PDP лежит идея оценки среднего значения прогноза модели для всех объектов, при фиксированных значениях определенных признаков, варьируя значения всех остальных признаков.

# Предлагаемый алгоритм автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения




Алгоритм автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения:

1) создает и обучает модель на основе решающего дерева, которую в дальнейшем интерпретирует;
2) интерпретирует полученную модель глобально методами SHAP и PDP, затем визуализирует результаты интерпретации;
3) интерпретирует полученную модель локально методам LIME, затем визуализирует результаты интерпретации.

# Программная реализация

Для реализации предложенного алгоритма использовался язык программирования *Python*.
Для реализации предложенного алгоритма создан класс *Inter*, включающий в себя методы класса, реализующие поставленные задачи. В качестве входных параметров класс *Inter* принимает набор данных *data*, целевую переменную *metka*, номер объекта *idd*, который необходимо интерпретировать локально, тип решаемой задачи *model_type*. Метод класса *__init__* производит инициализацию входных параметров. Метод класса *model* создает и обучает модель машинного обучения с помощью модели случайного леса. Метод класса *inter_global* глобально интерпретирует обученную модель машинного обучения методами SHAP и PDP и затем визуализирует результаты интерпретации. Метод класса *inter_local* локально интерпретирует обученную модель машинного обучения методом LIME и затем визуализирует результаты интерпретации.

# Вычислительные эксперименты

Вычислительные эксперименты проводились на существующих наборах данных:

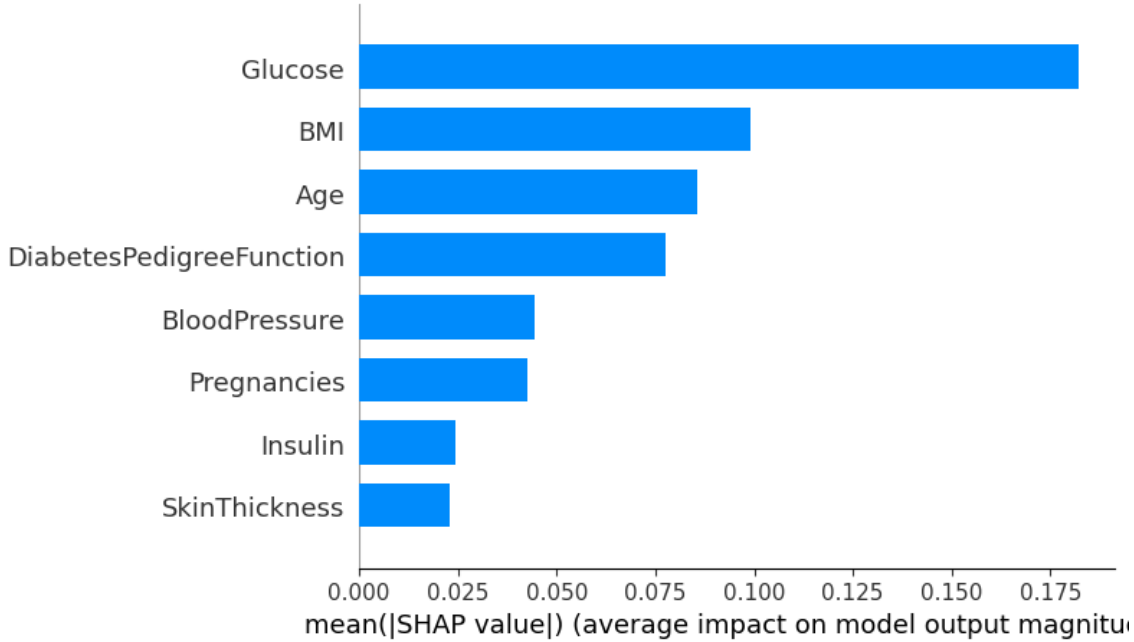
- ***Набор данных для эксперимента 1***

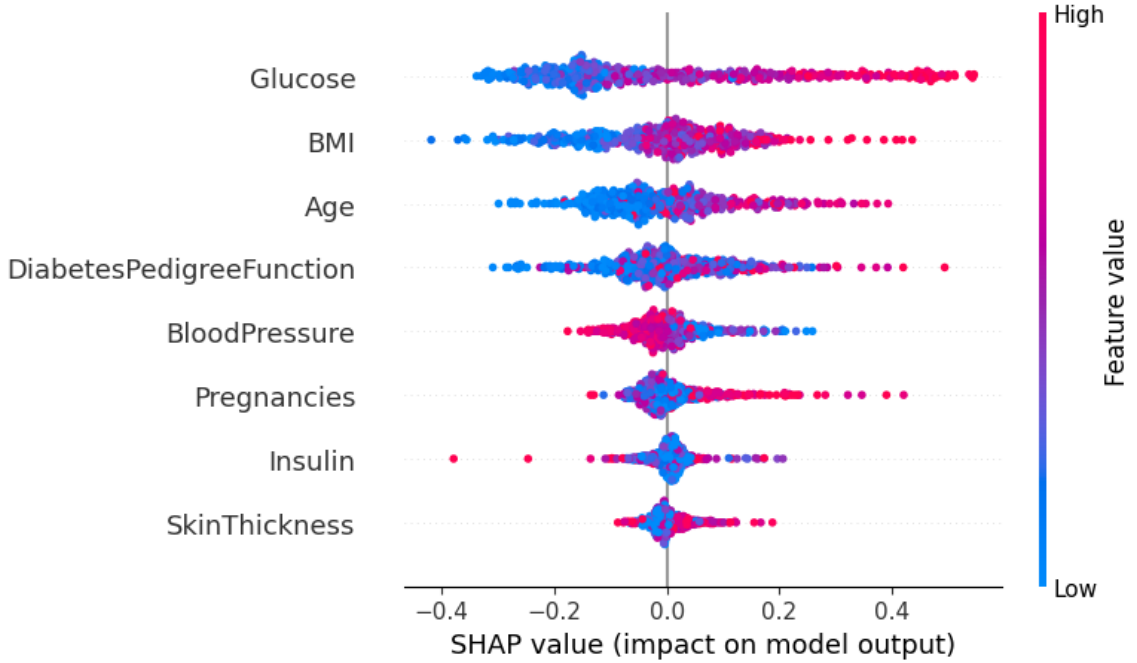
Набор данных ***diabetes***, который содержит медицинские данные, влияющие на риск развитие сахарного диабета. Объекты набора данных состоят из 8 признаков и целевой переменной: Pregnancies - количество беременностей; Glucose - плазменные концентрации глюкозы в крови; BloodPressure - диастолическое артериальное давление; SkinThickness - толщина кожи в области трицепса; Insulin - количество инсулина в крови; BMI - индекс массы тела; DiabetesPedigreeFunction - оценка предрасположенности к диабету; Age - возраст; Outcome – целевая переменная, показывающая прогноз, на предрасположенность к заболеванию сахарным диабетом в ближайшие пять лет. Набор данных состоит из 768 объектов.

- ***Набор данных для эксперимента 2***

Набор данных ***california housing***, который содержит данные о средней стоимость домов в Калифорнии в зависимости от квартала. Объекты набора данных состоят из 8 признаков и целевой переменной: Longitude - долгота квартала с недвижимостью; Latitude - широта квартала с недвижимостью; HouseAge - медиана возраста домов в квартале; AveRooms - общее количество комнат в квартале; AveBedrms - общее количество спален в квартале; Population - население квартала; AveOccup - количество семей в квартале; MedInc - медианный доход в квартале; MedHouseVal - целевая переменная, показывающая медианную стоимость дома в квартале. Набор данных состоит из 20640 объектов.

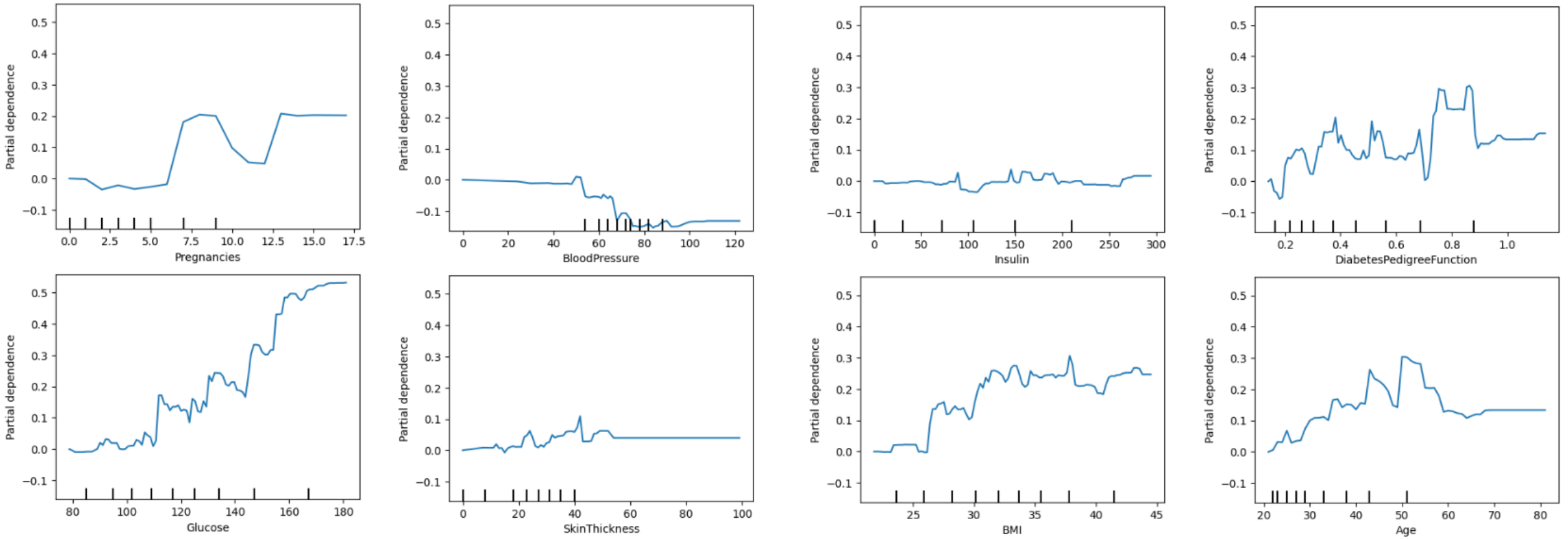
# Эксперимент 1. Метод SHAP



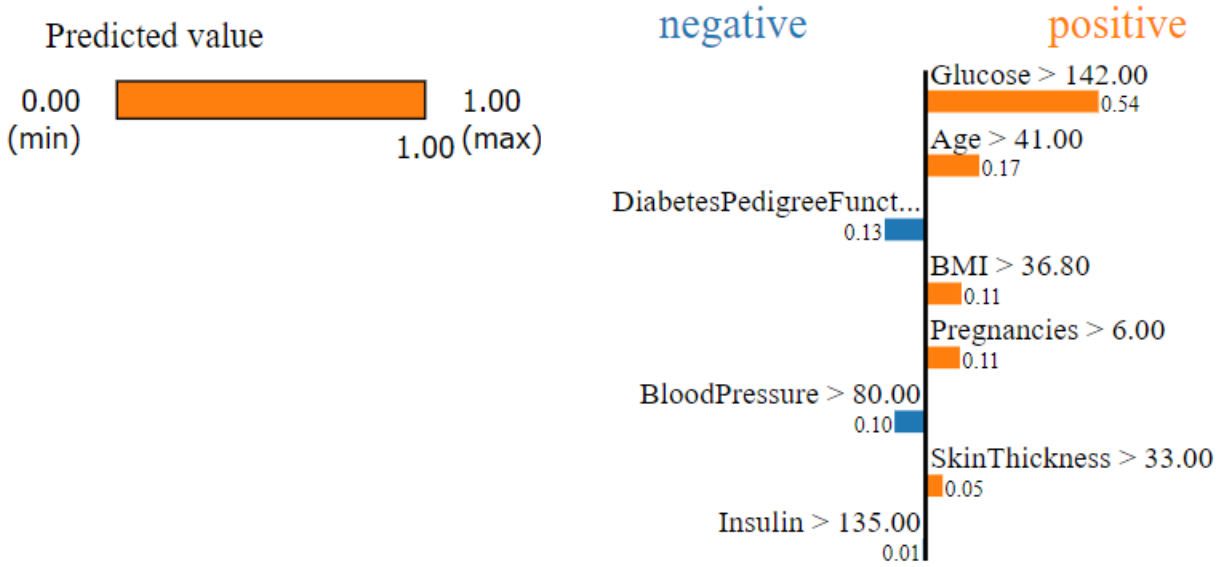

Из левого рисунка можно сделать вывод о важности признаков для модели: чем больше значение по оси абсцисс, тем более значимым является признак при оценке предрасположенности развития сахарного диабета в течении ближайших пяти лет.
Из правого рисунка можно сделать вывод о том какой вклад в оценку предрасположенности развития сахарного диабета в течении ближайших пяти лет вносят значения признаков: отрицательный вклад вносят значения расположенные в левой полуплоскости, положительный вклад вносят значения расположенные в правой полуплоскости, цвет точек отражает значение признака (чем краснее цвет, тем выше значение признака).

# Эксперимент 1. Метод PDP




Из рисунка можно сделать выводы аналогичные выводам, представленным на предыдущем слайде для левого рисунка.

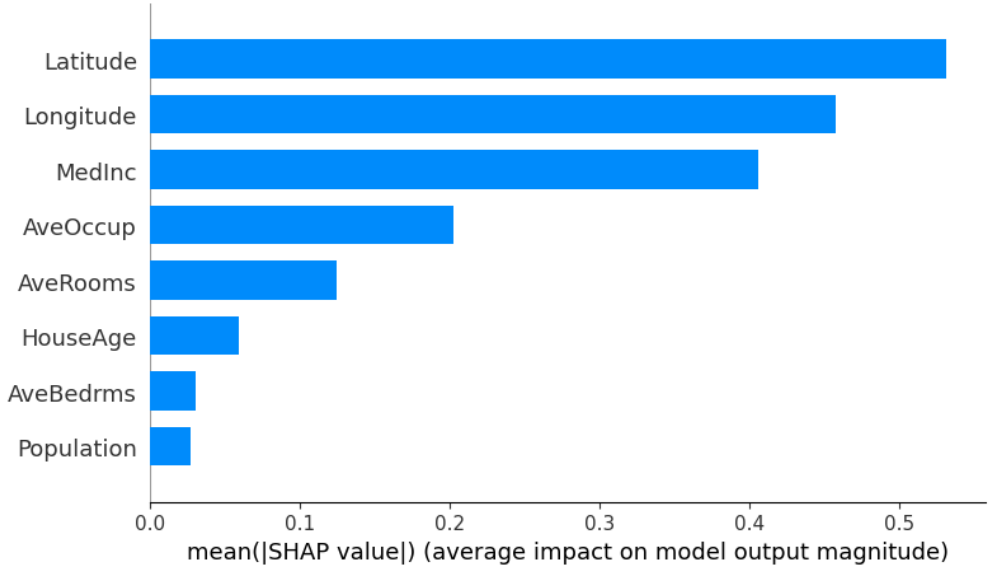
# Эксперимент 1. Метод LIME

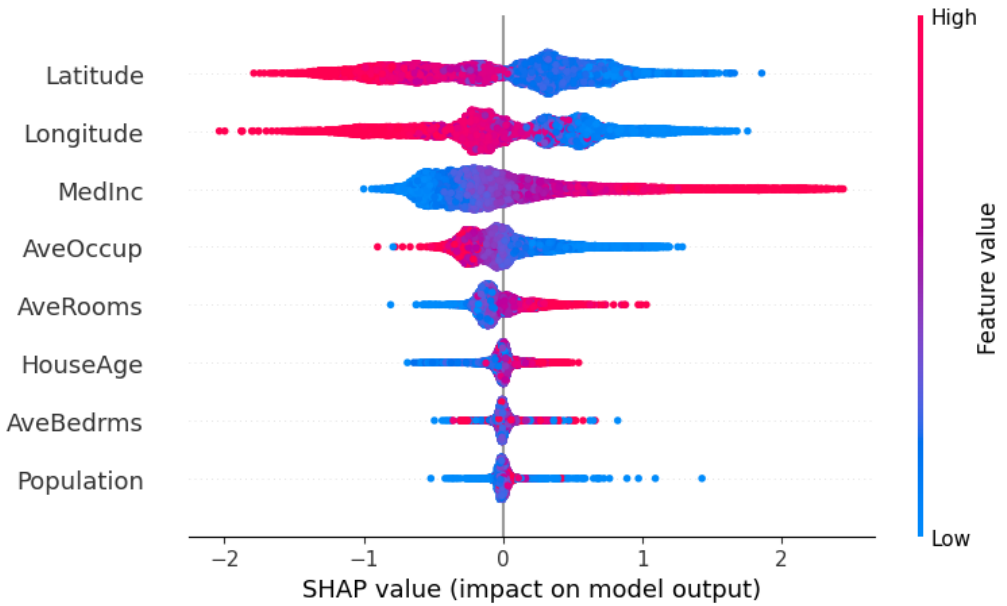



| Feature | Value |
|---|---|
| Glucose | 167.00 |
| Age | 43.00 |
| DiabetesPedigreeFunction | 0.17 |
| BMI | 37.60 |
| Pregnancies | 8.00 |
| BloodPressure | 106.00 |
| SkinThickness | 46.00 |
| Insulin | 231.00 |

Из рисунка можно сделать вывод о том какие признаки вносят вклад в риск развития сахарного диабета в течении ближайших пяти лет : признаки выделенные оранжевым цветом вносят положительный вклад, признаки выделенные синим цветом вносят отрицательный вклад.

# Эксперимент 2. Метод SHAP

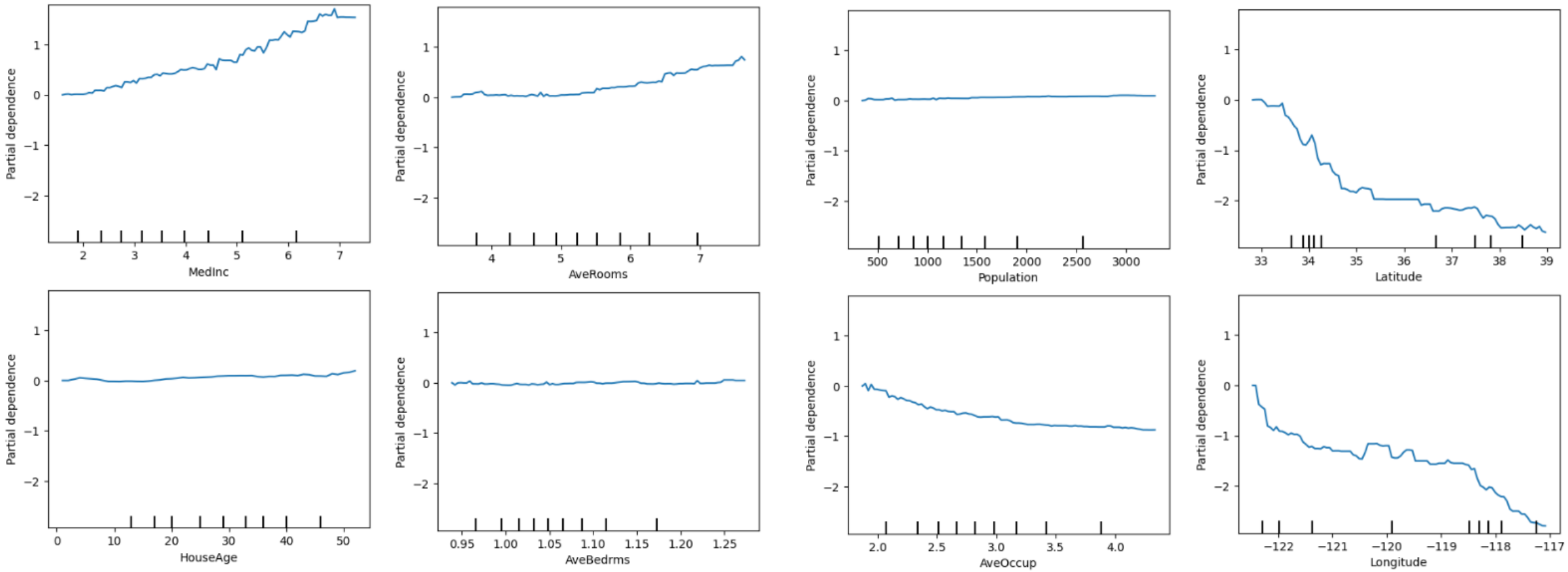



Из левого рисунка можно сделать вывод о важности признаков для модели: чем больше значение по оси абсцисс, тем более значимым является признак при оценке медианной стоимости дома в зависимости от квартала.

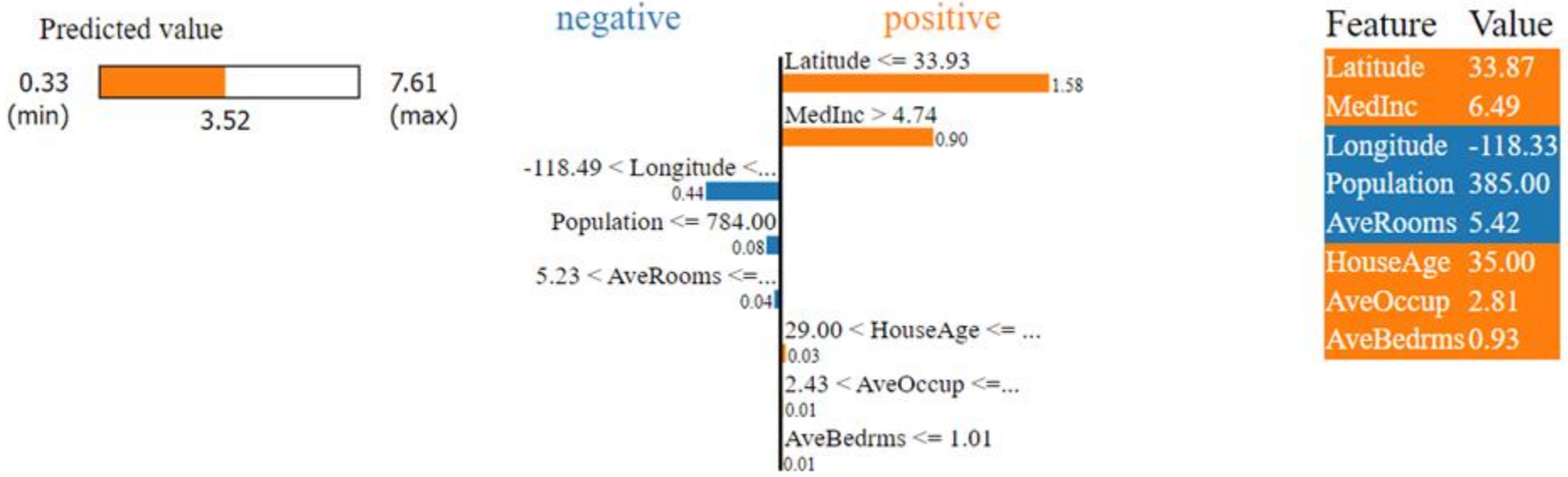
Из правого рисунка можно сделать вывод о том какой вклад в оценку медианной стоимости дома в зависимости от квартала вносят значения признаков: отрицательный вклад вносят значения расположенные в левой полуплоскости, положительный вклад вносят значения расположенные в правой полуплоскости, цвет точек отражает значение признака (чем краснее цвет, тем выше значение признака).

# Эксперимент 2. Метод PDP




Из рисунка можно сделать выводы аналогичные выводам, представленным на предыдущем слайде для левого рисунка

# Эксперимент 2. Метод LIME




Из рисунка можно сделать вывод о том какой вклад в медианную стоимость дома вносят признаки: признаки выделенные оранжевым цветом вносят положительный вклад, признаки выделенные синим цветом вносят отрицательный вклад.

# Заключение

В результате проведенной работы:

1) изучены аппроксимирующие модели машинного обучения;
2) изучены методы интерпретации аппроксимирующих моделей машинного обучения;
3) разработано ПО, реализующее предложенный алгоритм автоматизации процесса интерпретации;
4) проведены вычислительные эксперименты, показавшие эффективность разработанного ПО.

# БЛАГОДАРЮ ЗА ВНИМАНИЕ!