**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.
Консультант: к.т.н., доцент, Агасиев Т.А.
Научный руководитель: д.ф-м.н., профессор, Карпенко А.П.

# Введение

Интерпретация аппроксимирующих моделей машинного обучения - это процесс понимания того, как модель принимает решения и какие признаки оказывают наибольшее влияние на результат.

# Цели и задачи

Цель работы – разработка ПО для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения.

Задачи:

- Изучить аппроксимирующие модели машинного обучения;
- Изучить методы интерпретации моделей машинного обучения;
- Изучить существующее ПО для интерпретации;
- Реализовать ПО для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения;
- Анализ работы реализованного ПО.

# 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

# Решающее дерево

Рассмотрим решающее дерево, в котором:

- каждой внутренней вершине $v$ приписан предикат $B_v$;
- каждой листовой вершине $v$ приписан прогноз $c_v \in Y$, где $Y$ — область значений целевой переменной (в случае классификации листу может быть также приписан вектор вероятностей классов).

Предикат $B_v$ может иметь, произвольную структуру, но, как правило, на практике используют сравнение с порогом $t \in R$ по произвольному $j$-му признаку:

$$B_v(x, j, t) = [x_j \leq t].$$

При проходе через узел дерева с данным предикатом объекты будут отправлены в правое поддерево, если значение $j$-го признака у них меньше либо равно $t$, и в левое — если больше.

В ходе предсказания осуществляется проход по этому дереву к некоторому листу. Для каждого объекта выборки $x$ движение начинается из корня. В очередной внутренней вершине $v$ проход продолжится вправо, если $B_v(x)$ меньше порогового значения $t$, и влево, если $B_v(x)$ больше или равен порогового значения $t$. Проход продолжается до момента, пока не будет достигнут некоторый лист, и ответом алгоритма на объекте $x$ считается прогноз $c_v$, приписанный этому листу.

# Методы интерпретации моделей машинного обучения

Локальные:

- LIME
- SHAP
- ICE

Глобальные:

- PDP
- ALE

# Метод LIME

Суть метода LIME заключается в создании локальной модели, которая объясняет прогноз модели на конкретном объекте.

Математическое описание метода LIME:

$$E(x) = arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Модель объяснения для объекта $x$ – это локальная модель $g$, которая минимизирует функцию потерь $L$, которая измеряет, насколько близким является объяснение к прогнозу исходной модели машинного обучения $f$, при этом сохраняя низкую сложность модели $\Omega(g)$. $G$ - это семейство возможных локальных моделей. Мера близости $\pi_x$ определяет размер окрестности вокруг объекта $x$, которая рассматривается для объяснения.

# Метод SHAP

Суть метода SHAP заключается в вычислении значений Шепли для каждого признака и объединении их в одну величину, которая показывает важность признака для прогноза модели на конкретном объекте.
Математическое описание метода SHAP:

$$\Delta_f(i,S) = E[f(x)|x_{S\cup i}] - E[f(x)|x_S],$$

здесь $x_S$ - признаки, для которых должна быть рассчитаны значения SHAP, $x_{S\cup i}$ - признаки, стоящие перед $x_S$, $\Delta(i,S)$ - изменение в предсказании $x$ между условным математическим ожиданием $E[f(x)|x_S]$ признака, для которого рассчитываются значения SHAP, и условным математических ожиданием $E[f(x)|x_{S\cup i}]$ признаков, стоящих перед $x_S$.

# Метод PDP

В основе метода PDP лежит идея оценки среднего значения прогноза модели для всех объектов, при фиксированных значениях определенных признаков, варьируя значения всех остальных признаков.

Математическое описание метода PDP:

Частичная функция $f_S$ оценивается путем расчета средних значений на тренировочных входных данных, также известный как метод Монте-Карло:

$$f_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)}),$$

здесь $x_S$ представляют собой признаки, для которых должна быть построена частичная функция зависимости, $x_C^{(i)}$ представляют собой значения признаков из набора данных, которые не рассматриваются, $n$ - количество объектов в наборе данных.

Важность определяется отклонением каждого уникального значения признака от средней кривой:

$$I(x_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(f_S\left(x_S^{(k)}\right) - \frac{1}{K}\sum_{k=1}^{K} f_S(x_S^{(k)}))}^{2},$$

здесь $x_S^{(k)}$ представляет собой $K$ уникальных значений признака $X_S$.

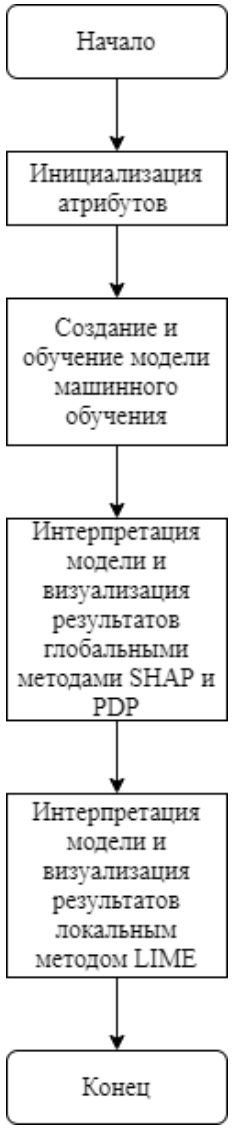
# Алгоритм автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения

Программная реализация ПО, автоматизирующего интерпретацию моделей машинного обучения, включает в себя следующий функционал:

1) создание и обучение модели машинного обучения, которую необходимо интерпретировать;
2) интерпретация модели машинного обучения различными методами;
3) визуализация интерпретации модели машинного обучения.

Для реализации ПО 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 на основерешающих деревьев. Метод класса *inter_global* глобально интерпретирует модель машинного обучения методами SHAP и PDP и затем визуализирует результаты интерпретации. Метод класса *inter_local* локально интерпретирует модель машинного обучения методом LIME и затем визуализирует результаты интерпретации.

# Блок-схема разработанного алгоритма

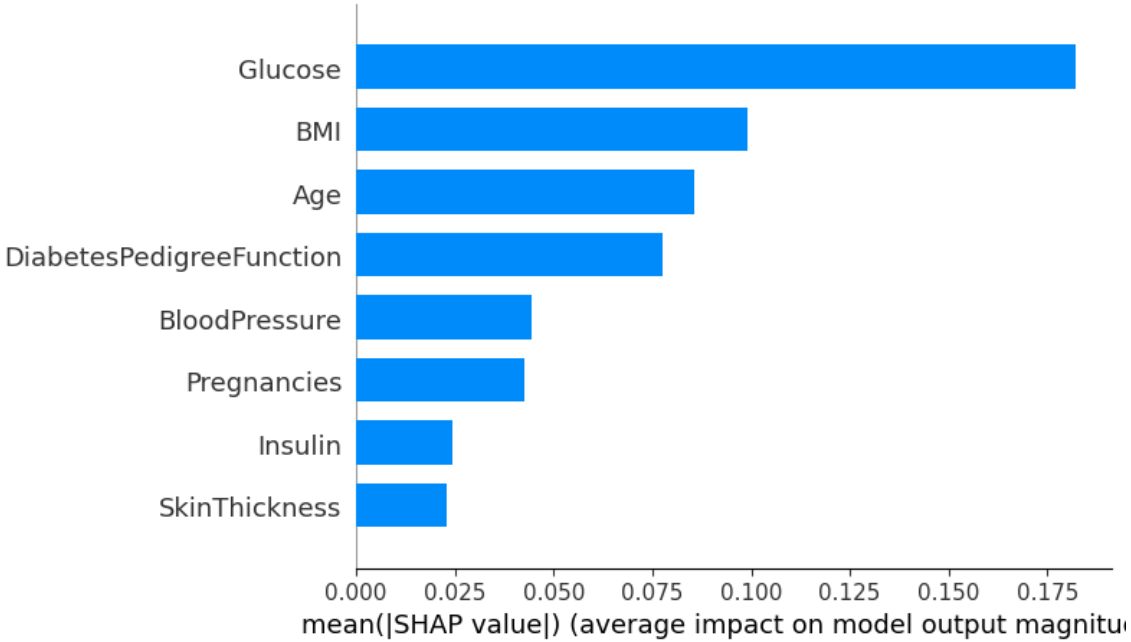


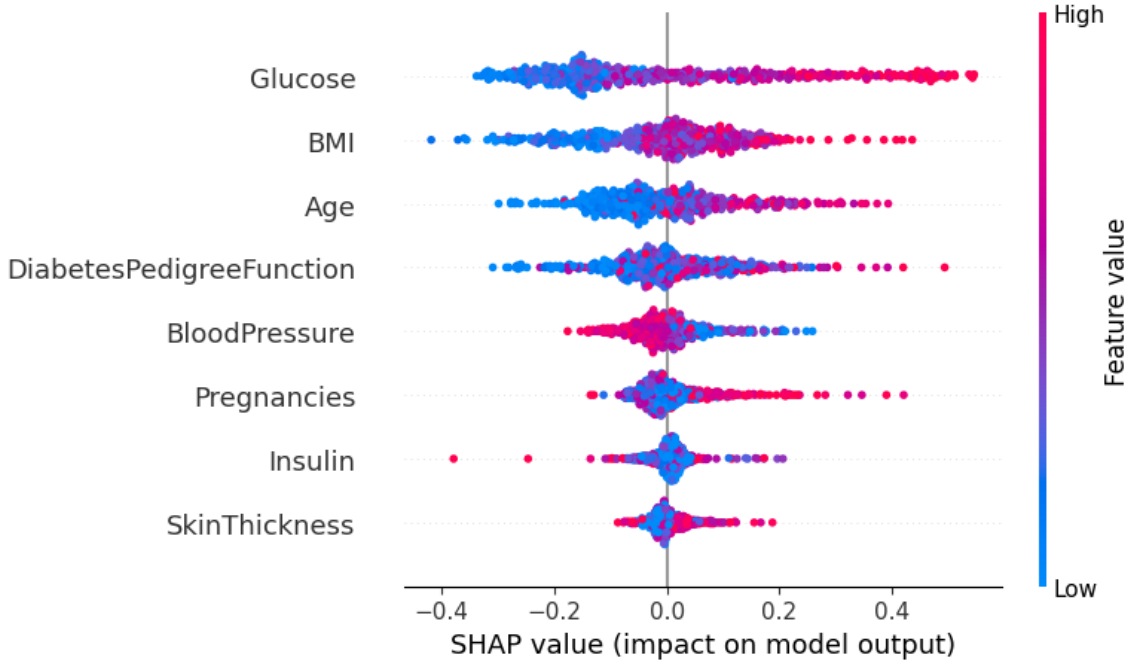
# Вычислительные эксперименты

Вычислительные эксперименты проводились на существующих наборах данных:

- ***diabetes***, который содержит медицинские данные, влияющие на риск развитие сахарного диабета. Набор данных состоит 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.
- ***california housing***, который содержит данные о средней стоимость домов в Калифорнии в зависимости от квартала. Набор данных состоит 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.

# Эксперимент 1

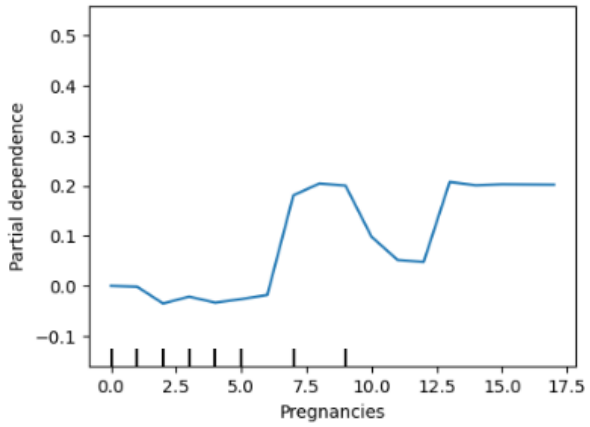


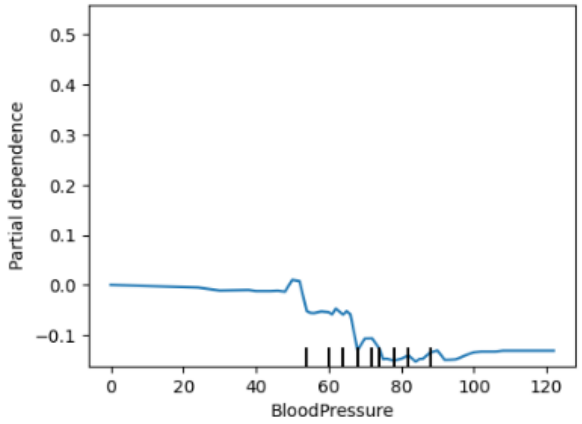



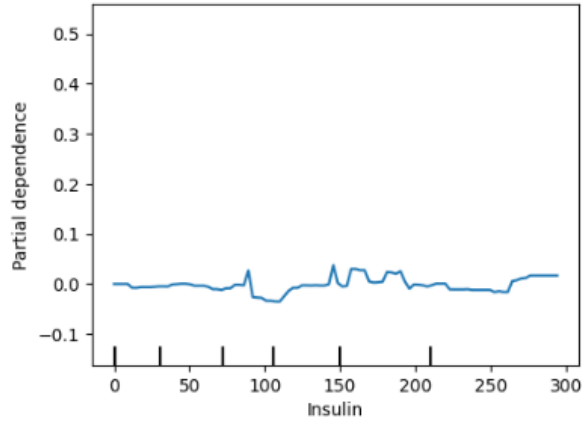

Из левого рисунка можно сделать вывод, что чем выше значение таких признаков как плазменные концентрации глюкозы в крови, индекс массы тела, возраст, оценка предрасположенности к диабету, количество беременностей и толщина кожи в области трицепса, тем выше риск развития сахарного диабета в течении пяти лет. При этом, чем выше значения признаков количество инсулина в крови и диастолическое артериальное давление, тем риск развития сахарного диабета в течении пяти лет ниже.

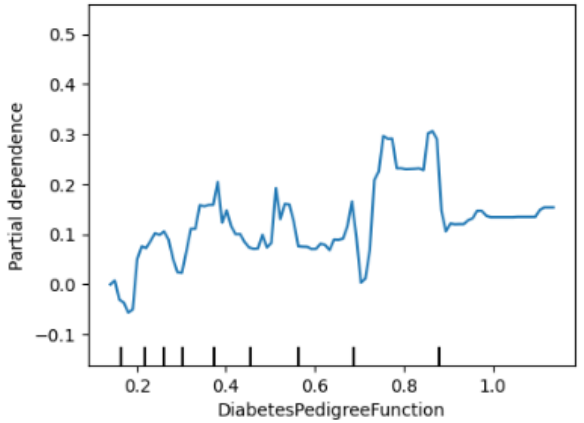
Из правого рисунка можно сделать вывод, что самыми значимыми признаками, влияющими на развитие сахарного диабета в ближайшие пять лет, являются плазменные концентрации глюкозы в крови, индекс массы тела и возраст. Наименее значимыми признаками являются количество инсулина в крови и толщина кожи в области трицепса. Признаки оценка предрасположенности к диабету, диастолическое артериальное давление, количество инсулина в крови вносят умеренный вклад в риск развития сахарного диабета в ближайшие пять лет у объекта.

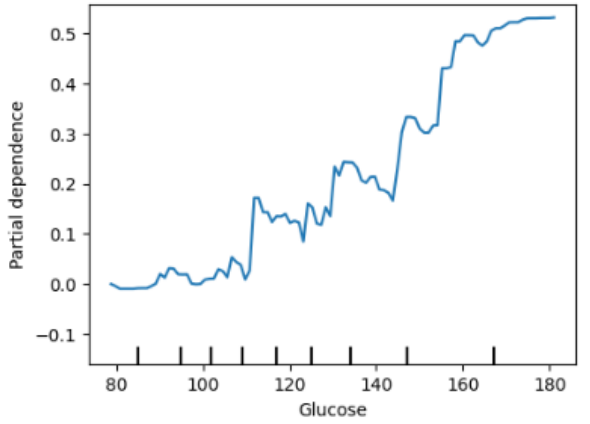
# Эксперимент 1

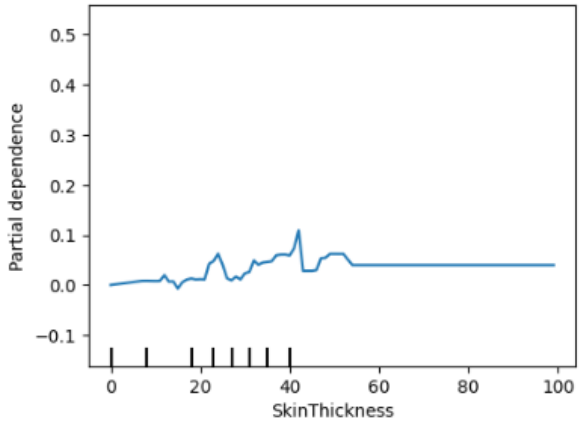


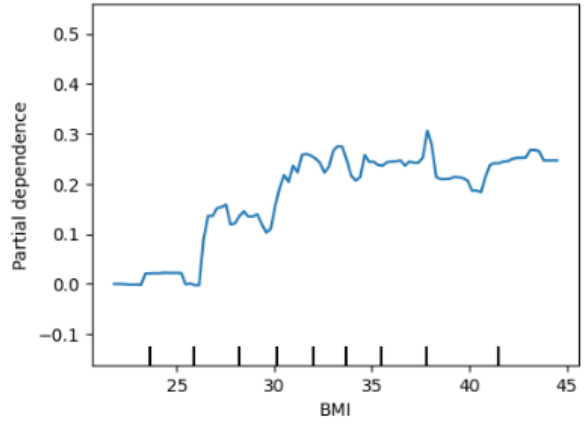



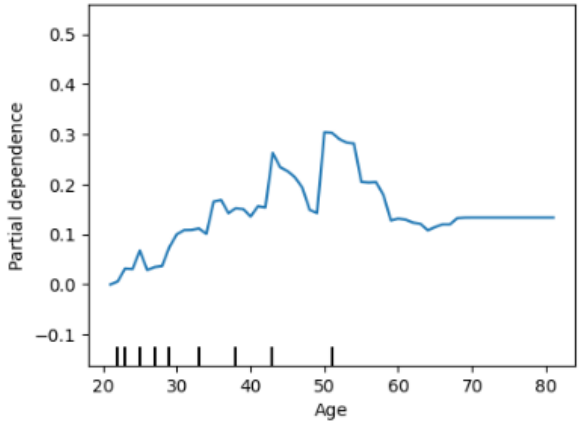



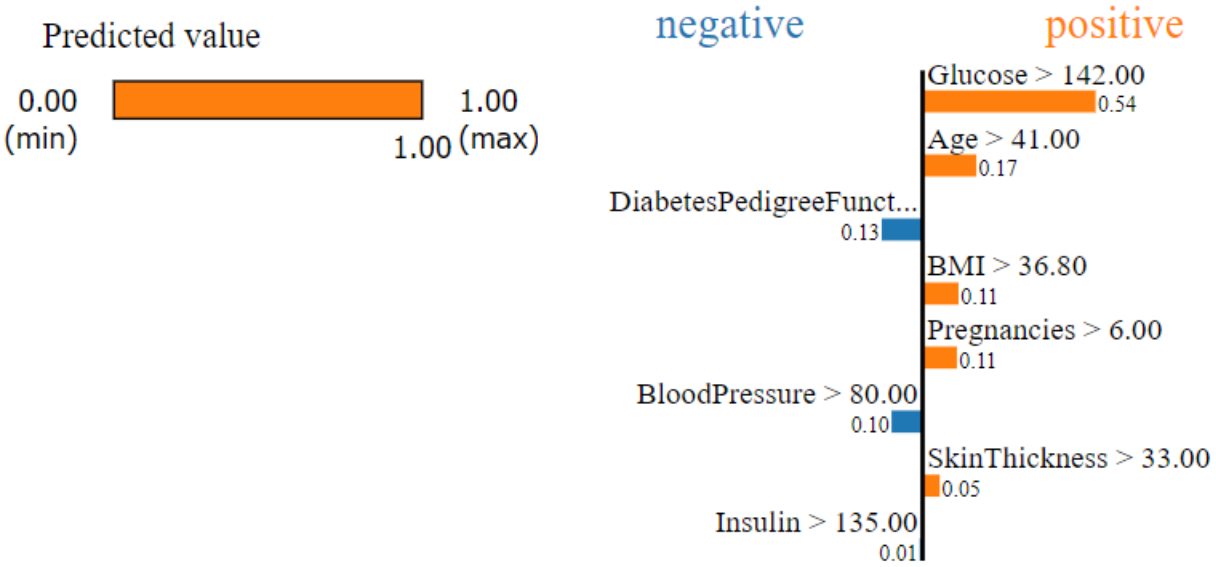




Из рисунка можно сделать выводы аналогичные выводам для левого рисунка представленным на предыдущем слайде

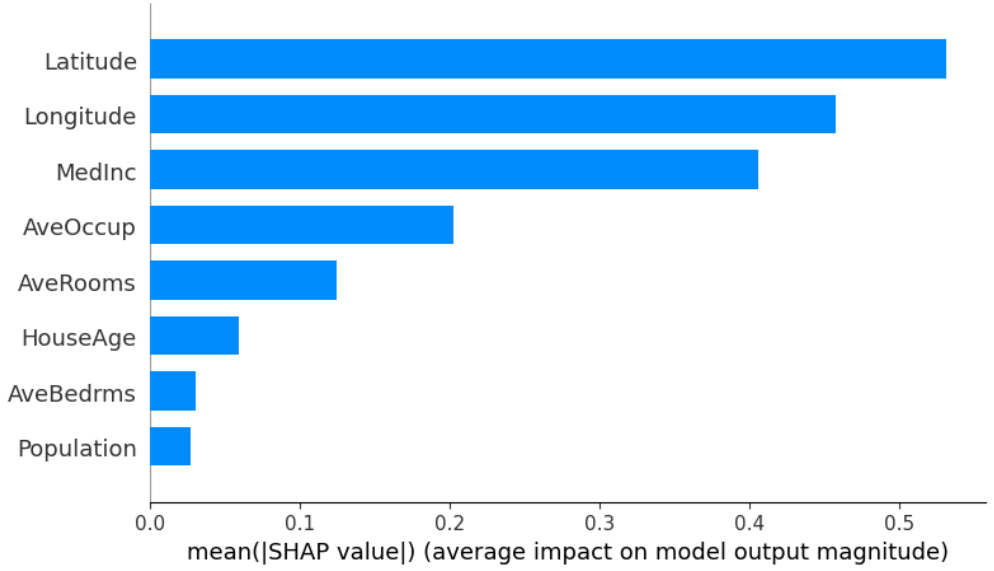
# Эксперимент 1

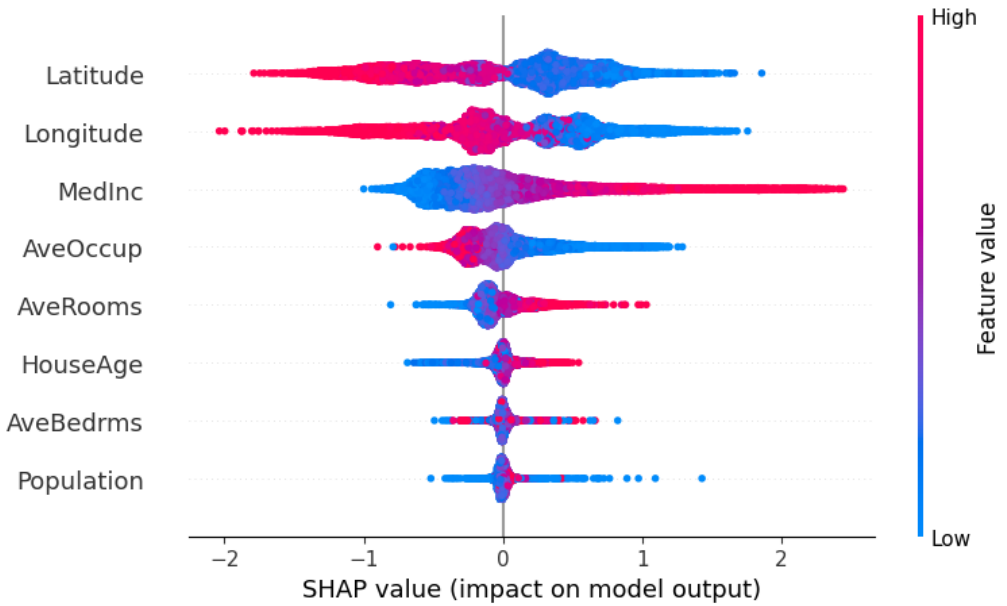



| Feature | Value |
|---|---|
| Glucose | 167.00 |
| Age | 43.00 |
| DiabetesPedigreeFunction | 0.17 |
| BMI | 37.60 |
| Pregnancies | 8.00 |
| BloodPressure | 106.00 |
| SkinThickness | 46.00 |
| Insulin | 231.00 |

Из рисунка можно сделать вывод, что несмотря на то что значения признаков оценка предрасположенности к диабету, диастолическое артериальное давление и количество инсулина в крови не приводят к риску развития сахарного диабета у объекта, значения остальных признаков показывают, что у объекта есть риск развития сахарного диабета в течении пяти лет.

# Эксперимент 2

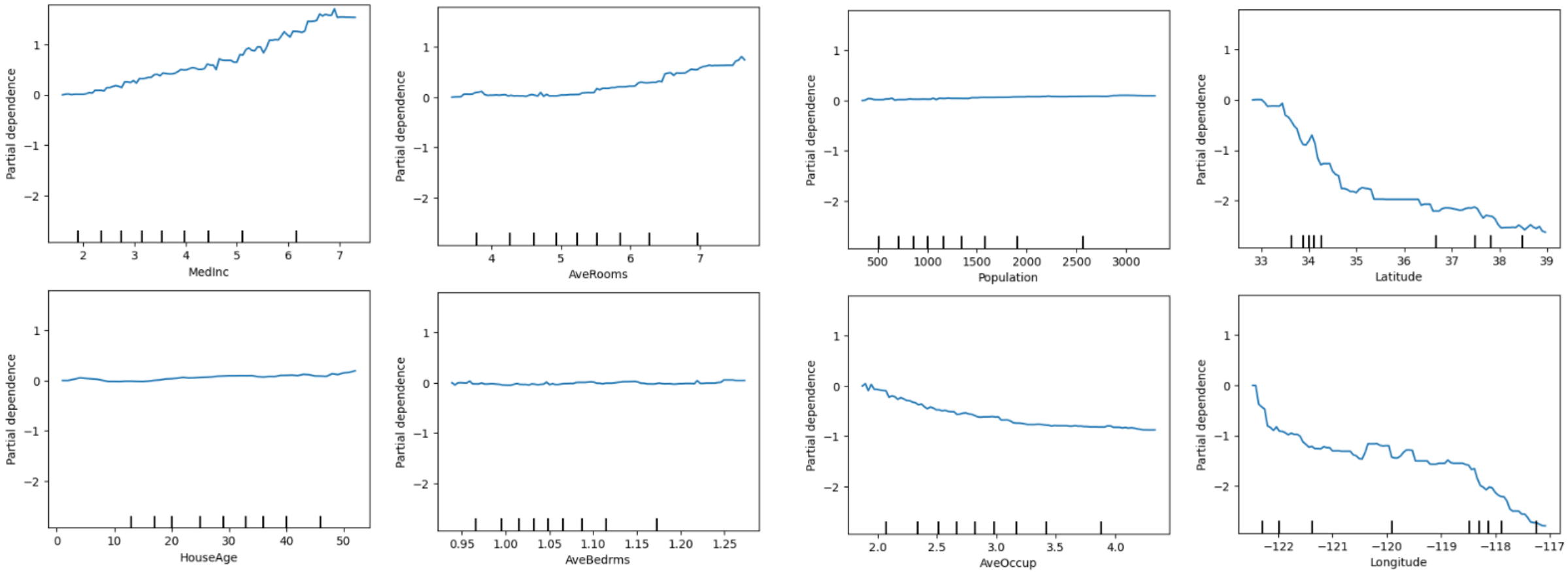



Из левого рисунка можно сделать вывод, что самыми значимыми признаками, влияющими медианную стоимость дома в квартале, являются широта квартала с недвижимостью, долгота квартала с недвижимостью, медианный доход в квартале и количество семей в квартале. Наименее значимыми признаками являются общее количество спален в квартале и население квартала. Признаки общее количество комнат в квартале и медиана возраста домов в квартале вносят умеренный вклад в медианную стоимость дома в квартале.

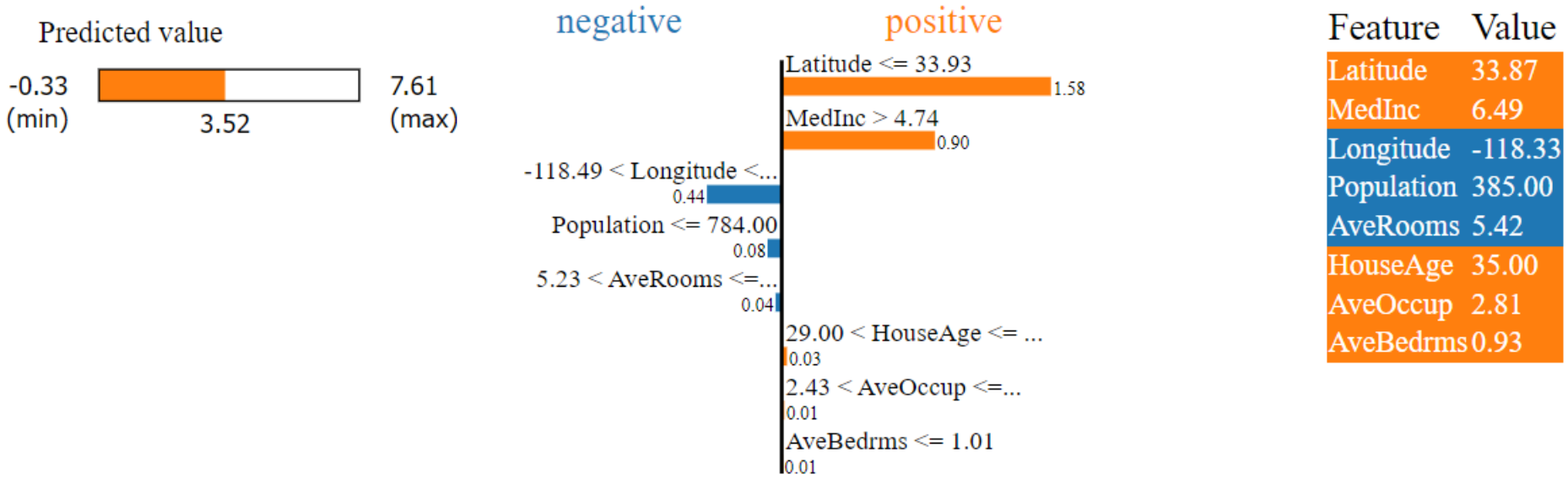
Из правого рисунка можно сделать вывод, что чем выше значение таких признаков как общее количество комнат в квартале, медиана возраста домов в квартале, медианный доход в квартале тем выше медианная стоимость дома в квартале. При этом, чем выше значения признаков широта квартала с недвижимостью, долгота квартала с недвижимостью, количество семей в квартале, тем ниже медианная стоимость дома в квартале. Значения признаков количество спален в квартале и население квартала почти не влияют на медианная стоимость дома в квартале.

# Эксперимент 2




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

# Эксперимент 2




Из рисунка можно сделать вывод, что на повышение медианной стоимости дома в квартале повлияли все признаки кроме долготы квартала с недвижимостью, общего количество комнат в квартале и населения квартала. Они снизили медианную стоимость дома в квартале.

# Заключение

В результате проведенной работы:

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

БЛАГОДАРЮ ЗА ВНИМАНИЕ!