

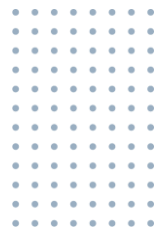


Казанский
Государственный
Медицинский
Университет

Деревья решений и случайный лес

Абзалилова Лия Рашитовна

Казань, 2025



План лекции



Деревья решений



Алгоритм CART



Алгоритм Random Forest



Важность признаков

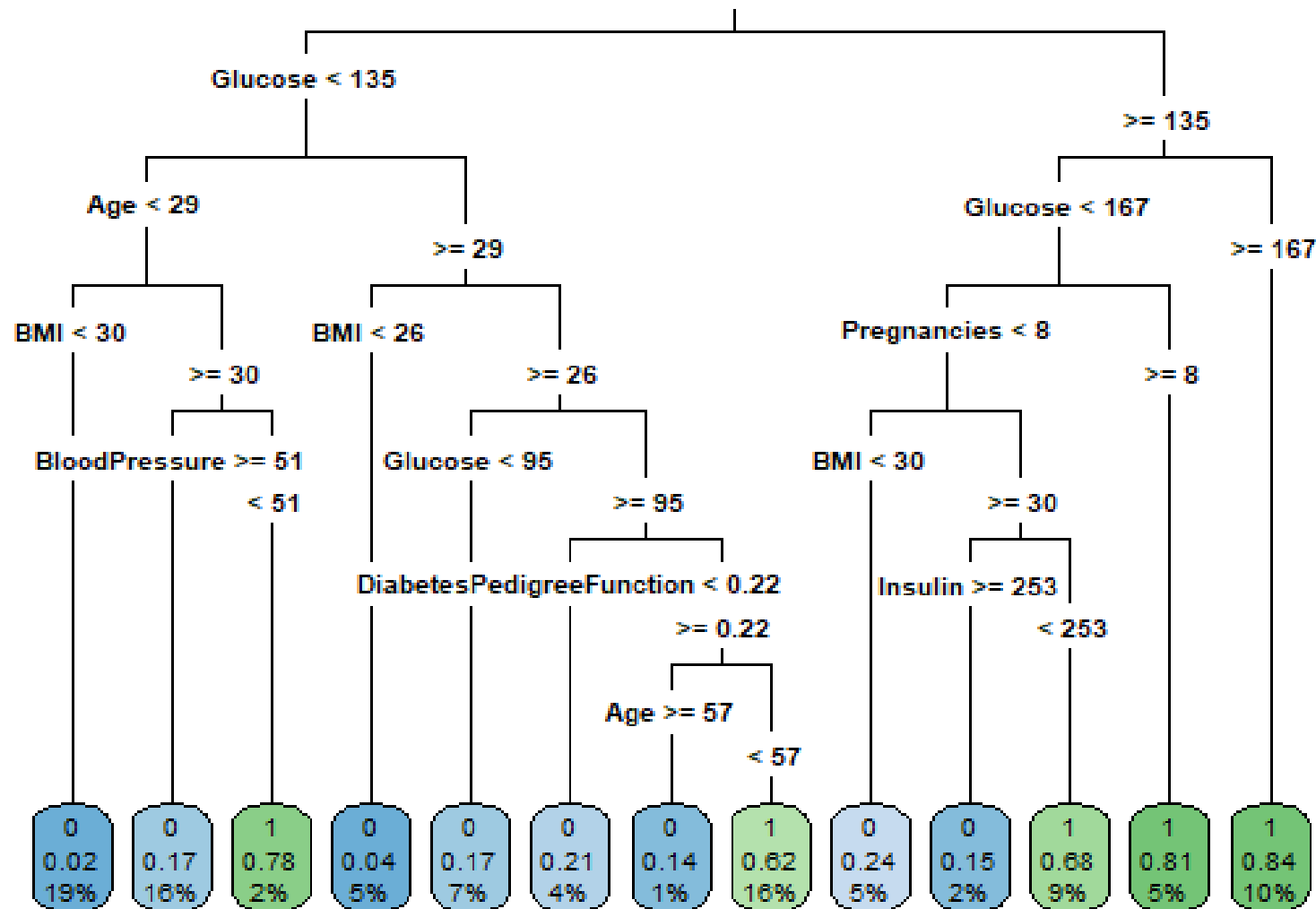
Деревья решений

Преимущества линейных методов:

- быстрота обучения;
- способность работать с большим количеством объектов и признаков;
- небольшое количество параметров;
- легкость регуляризации.

Недостаток — они могут восстанавливать только линейные зависимости

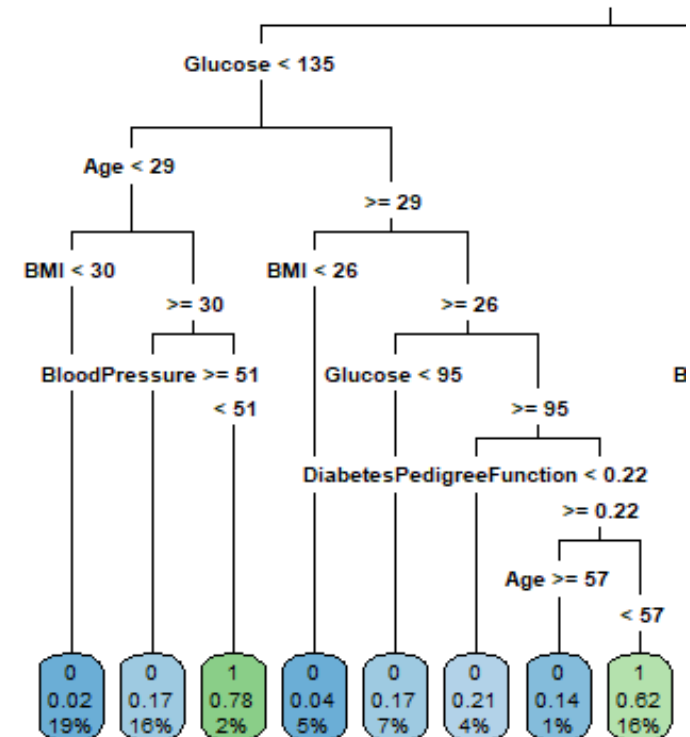
Пример



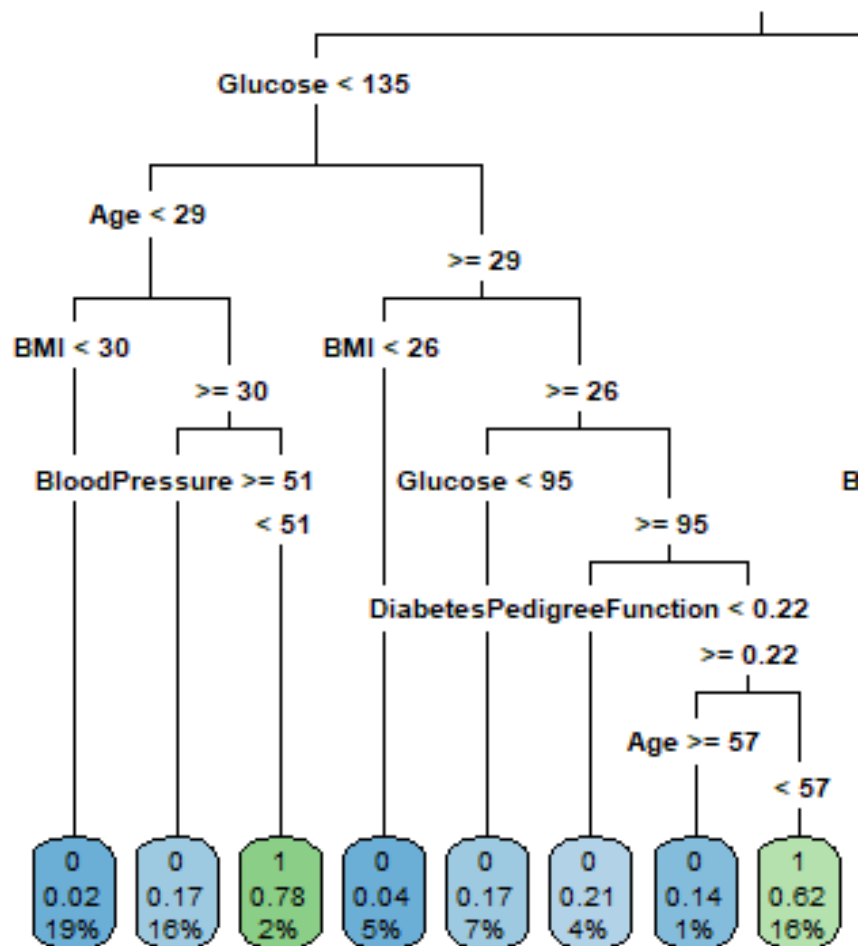
1. Число случаев беременности;
2. Концентрация глюкозы;
3. BloodPressure - артериальное диастолическое давление, мм. рт. ст.;
4. 2-х часовой сывороточный инсулин;
5. Индекс массы тела;
6. Числовой параметр наследственности диабета;
7. Возраст, лет;
8. Зависимая переменная (1 — наличие заболевания, 0 — отсутствие)

Деревья решений

- Деревья решений применяются для исследования взаимосвязи между результирующей переменной (target) и независимыми факторами (предикторами, регрессорами, features).
- Деревья решений применяются для решения задач регрессии и задач классификации.
- Дерево состоит из «листьев» и «веток»
- На «ветках» дерева имеются случаи, от которых зависит целевая функция (target), в «листьях» записаны значения целевой функции.
- Для классификации нового случая, следует спуститься по дереву до листа и выдать соответствующее значение target.



Принцип работы алгоритма «дерево решений»



- Весь набор данных называется корневым узлом
- Данные разбиваются на два или несколько узлов, так чтобы данные в узлах сильно отличались друг от друга. В качестве правил, позволяющих оценить различия в узлах, являются значения признаков (например, возраст, пол и др.)
- Качество разбиения контролируется статистическими критериями

CART

- Зависимая переменная и предикторы могут быть категориальными или количественными.
- Является бинарным деревом - каждый узел дерева при разбиении имеет только двух потомков: в одном правило выполняется, в другом правило не выполняется.
- Для построения дерева метод CART использует меры неоднородности. Эти меры используют принцип уменьшения неоднородности в узле.
- Деление узла происходит так, чтобы узлы-потомки стали более однородными, чем его узел-родитель. В абсолютно однородном узле (чистом) все наблюдения имеют одно и то же значение зависимой переменной (все объекты принадлежат к одной и той же категории зависимой переменной).

CART

Для оценки неопределенности в каждом узле используется мера неоднородности, например, индекс Джини:

$$G = 1 - 2 \sum_{i=1}^k d_{xi} d_{xi}^H + \sum_{i=1}^k d_{xi} d_{yi}$$

Где: d_{yi} – доля i -ой группы в общем объеме совокупности;

d_{xi} – доля i -ой группы в общем объеме признака;

d_{xi}^H – накопленная доля i -ой группы в общем объеме признака.

Если вся выборка в узле делится на 2 части, то коэффициент Джини рассчитывается как средневзвешенное на объем каждой из подвыборки.

Наилучшим считается такое разбиение, для которого индекс Джини минимален.

Проблема переобучения

Решение проблемы - прунинг:

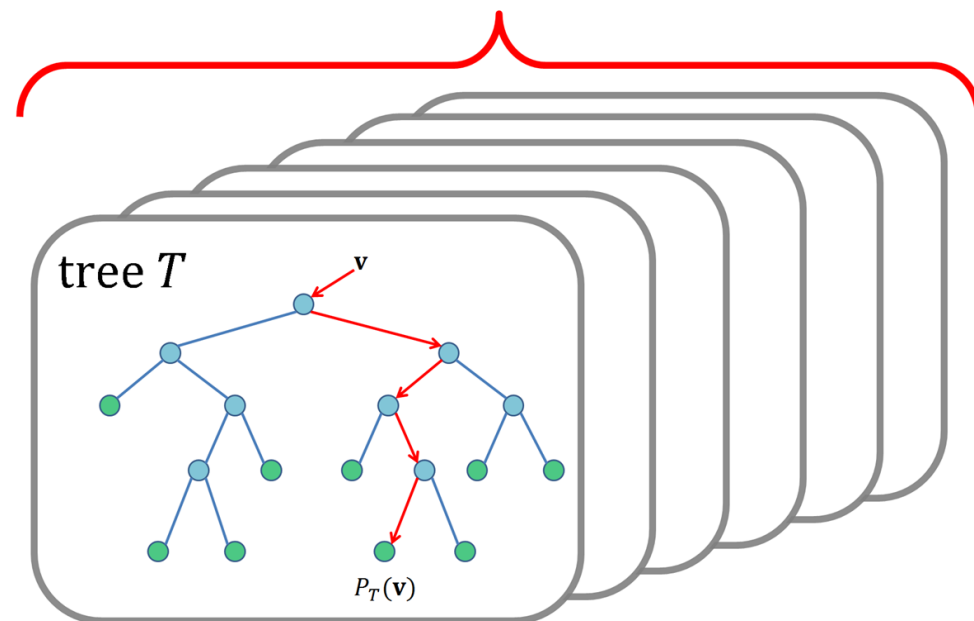
➤ контролировать глубину дерева, за счет удаления веток, дающих малый вес

Метод чувствителен к категориальным переменным, с большим количеством альтернатив:

➤ качество контролируется за счет перекрестной проверки (кросс-валидации)

➤ использовать ансамбль решающих деревьев.

Decision Forest



Алгоритм Random Forest

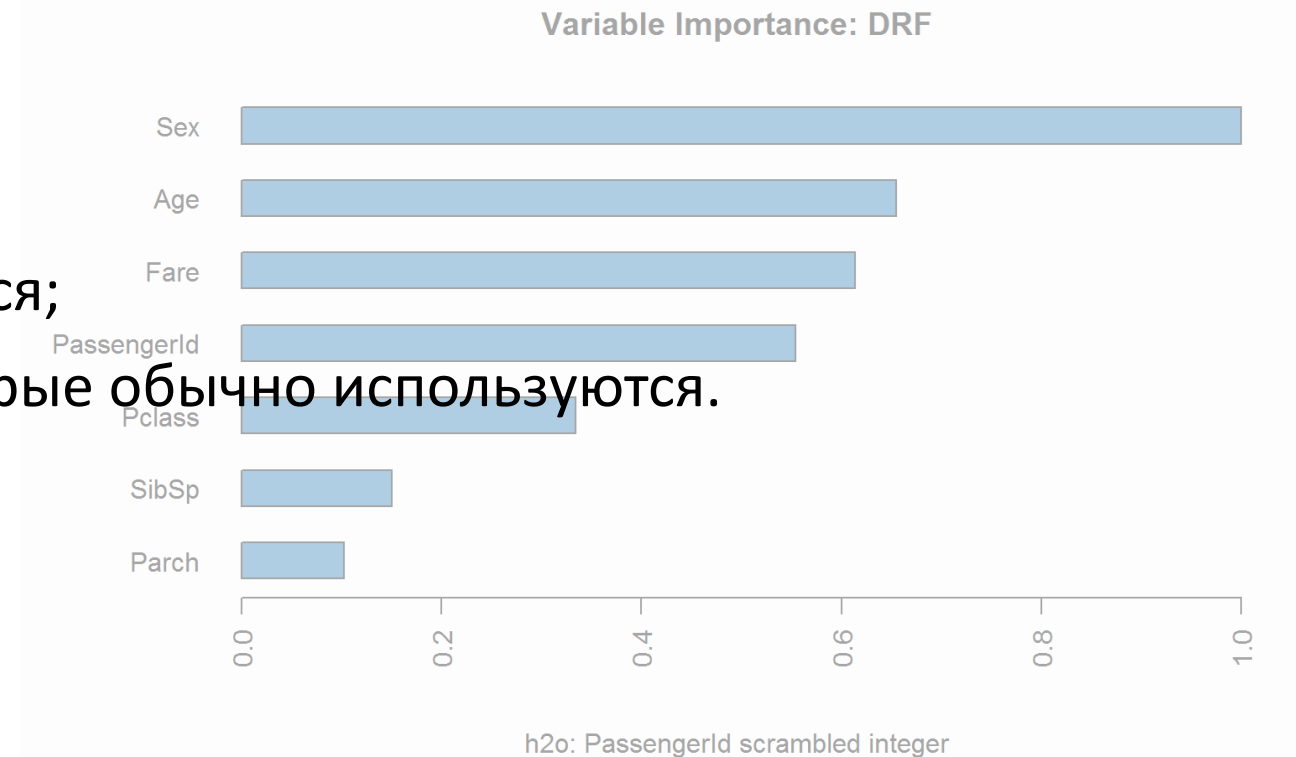
1. Создается случайная подвыборка с повторениями заданным размером (N) из обучающей выборки. *Будут образцы, которые в выборку не попадут.*
2. Строится решающее дерево, которое классифицирует все случаи данной подвыборки. При создании нового узла дерева выбирается случайный набор m признаков, на основе которых производится разбиение (то есть не все подряд выбираются, а отбирается лишь случайная часть из них). Выбор наилучшего из этих m признаков может осуществляться различными способами. Например, на основе критерия Джини.
3. Дерево строится до полного исчерпания подвыборки и не подвергается прунингу.
4. Классификация случаев проводится путём голосования: каждое дерево ансамбля относит классифицируемый объект к одному из классов, и побеждает тот класс, за который проголосовало наибольшее число деревьев.

Важность признака

Permutation Feature Importance (PFI) — это прием, который используется для объяснения моделей классификации и регрессии и основан на работе Бреймана Random Forests (Случайные леса). Общий принцип работы такой: метод случайным образом перетасовывает данные по одному признаку для всего набора данных и вычисляет показатель изменений интересующей метрики. Чем больше изменение, тем важнее компонент.

Permutation importance отличается:

- Его быстро посчитать;
- Широко применяется и легко понимается;
- Используется вместе с метриками, которые обычно используются.



Важность признака

Процесс выявления важности признаков выглядит следующим образом:

1. Получаем обученную модель на «нормальных» данных, вычисляем для нее метрики, в том числе и значение функции потерь.
2. Переставляем значения в одном столбце, прогнозируем с использованием полученного набора данных. Используем эти прогнозы и истинные целевые значения, чтобы вычислить, насколько функция потерь ухудшилась от перетасовки. Это ухудшение производительности измеряет важность переменной, которую только что перемешали.
3. Возвращаем данные к исходным значениям и повторяем шаг 2 со следующим столбцом в наборе данных, пока вычисляется важность каждого столбца.





**Спасибо за
внимание!**