Министерство науки и высшего образования Российской Федерации

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

Кафедра прикладной информатики

**Отчет по лабораторной работе №3 по дисциплине «Интеллектуальные системы»**

Выполнил

студент группы № 1462

\_\_\_\_\_\_\_\_\_\_\_\_А.В. Шабля

Проверил

доцент, канд. техн. наук

\_\_\_\_\_\_\_\_\_\_\_\_С.В. Аксёнов

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Томск – 2020

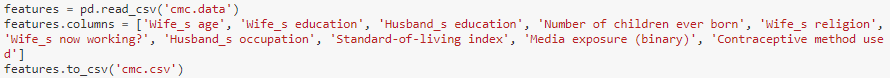
**Описание**

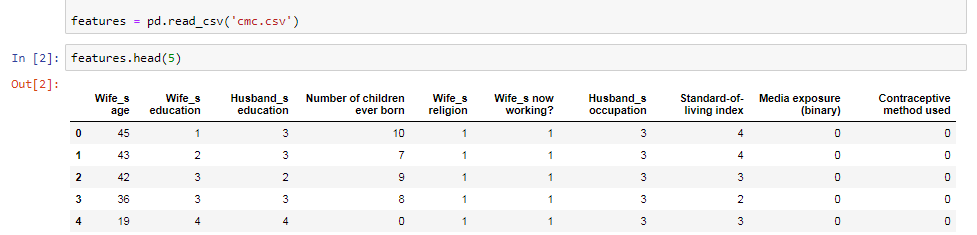
Данные содержат 1473 записей и 10 атрибутов. Набор данных является подмножеством Национального обследования распространенности контрацептивов в Индонезии 1987 года. Скачан с <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>.

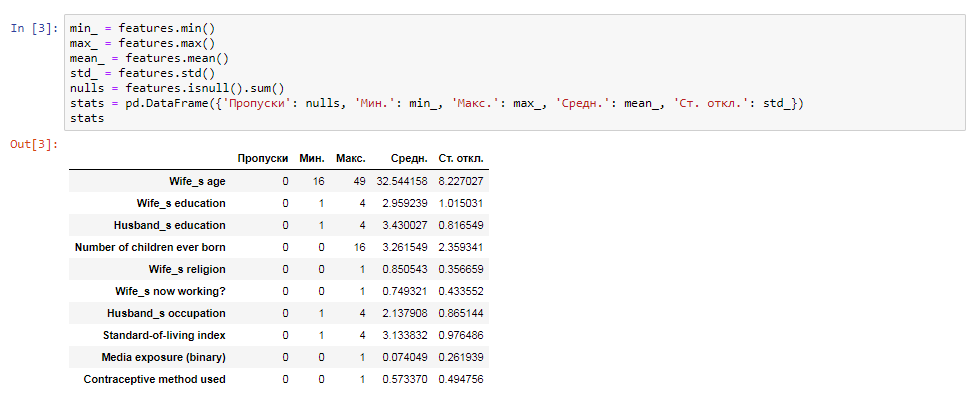
Attribute Information:

* Wife's age (numerical)
* Wife's education (categorical) 1=low, 2, 3, 4=high
* Husband's education (categorical) 1=low, 2, 3, 4=high
* Number of children ever born (numerical)
* Wife's religion (binary) 0=Non-Islam, 1=Islam
* Wife's now working? (binary) 0=Yes, 1=No
* Husband's occupation (categorical) 1, 2, 3, 4
* Standard-of-living index (categorical) 1=low, 2, 3, 4=high
* Media exposure (binary) 0=Good, 1=Not good
* Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term (будет изменено: 0=No-use, 1=use)

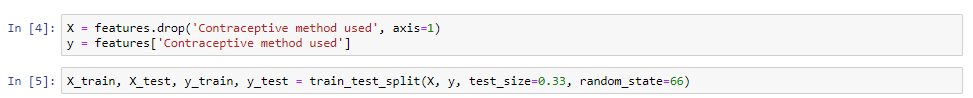
Исходная проблема состоит в том, чтобы предсказать текущий выбор метода контрацепции (без использования, долгосрочных или краткосрочных методов) женщины на основе ее демографических и социально-экономических характеристик. В данной лабораторной будут рассмотрены бинарные методы классификации, поэтому поля атрибута Contraceptive method used был изменен: то, что было «1» (No-use), стало «0», а то, что было «1» и «2» − было объединено в «1» (use). Для этого сначала был прочитан скачанный файл cmc.data, заданы наименования колонок, сохранен файл в csv, затем при помощи Excel’я совершены замены и уже по новой считан файл cmc.csv.







Разбиение данных на тренировочные и тестовые:

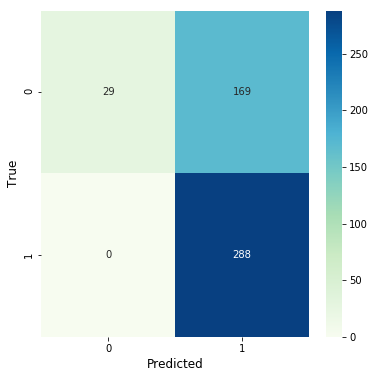


Матрица несоответствий - это просто квадратная таблица, которая сообщает о числе истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных. Хотя эти метрики можно легко вычислить вручную путем сравнения истинных и предсказанных меток классов, библиотека scikit-leaгn предлагает вспомогательную функцию confusion\_matrix предсказаний классификатора.

Точность (Precision) и полнота (Recall) - это метрики оценки качества, которые связаны с долями истинно положительных и истинно отрицательных исходов, при этом фактически полнота является синонимом доли истинно положительных исходов.

**Дерево решений**

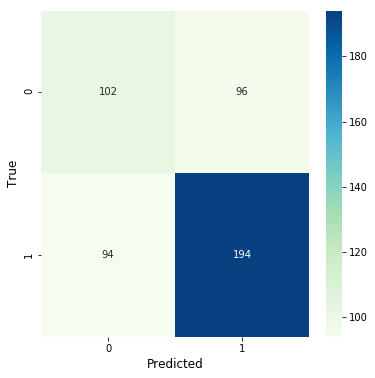
Деревья решений могут создавать сложные границы решения путем деления пространства признаков на прямоугольники. Однако мы должны быть осторожными, поскольку чем глубже дерево решений, тем сложнее становится граница решения, что может легко закончиться переобучением. Теперь, используя scikit-learn, натренируем дерево решений с максимальной глубиной 3, применив в качестве критерия неоднородности энтропию. Хотя для целей визуализации может понадобиться шкалирование признаков, отметим, что масштабирование признаков не является необходимой составной частью алгоритмов деревьев решений.



acc = 0.6522633744855967 | prec = 0.6301969365426696 | rec = 1.0 | f1 = 0.7731543624161074

**Случайный лес**

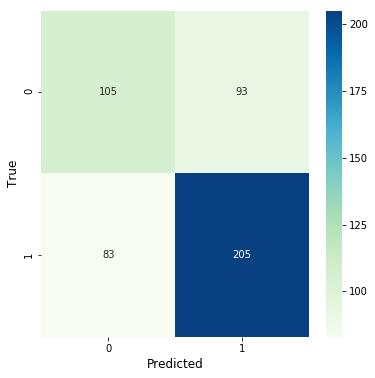
Интуитивно случайный лес можно рассматривать как ансамбль деревьев решений. В основе ансамблевого обучения лежит идея объединения слабых учеников для создания более устойчивой модели, т. е. сильного ученика, с более хорошей ошибкой обобщения и меньшей восприимчивостью к переобучению.



acc = 0.6090534979423868 | prec = 0.6689655172413793 | rec = 0.6736111111111112 | f1 = 0.671280276816609

**К-средних**

Основное преимущество такого подхода с запоминанием состоит в том, что классификатор немедленно адаптируется по мере сбора новых тренировочных данных. Однако его оборотная сторона - вычислительная сложность классифицирования новых образцов - растет линейно вместе с числом образцов в тренировочном наборе данных в наихудшем случае, если только в наборе данных не очень много размерностей (признаков) и алгоритм не был реализован с использованием эффективных структур данных, таких как КD-деревья. Кроме того, мы не можем отбросить тренировочные образцы, поскольку никакого трестирующего шага нет. Вследствие этого, если мы работаем с большими наборами данных, пространство памяти может представлять серьезную проблему.



acc = 0.6378600823045267 | prec = 0.6879194630872483 | rec = 0.7118055555555556 | f1 = 0.6996587030716723

**Вывод**

На практике были изучены три алгоритма классификации.

Мы узнали о разнообразных алгоритмах машинного обучения, которые используются для решения линейных и нелинейных задач. Мы убедились, что деревья решений особенно привлекательны, в случае если интерпретируемости модели уделено должное внимание.

Чтобы измерить общий успех, была использована функция F1, которая позволяет найти баланс между Precision и Recall.

* DecisionTreeClassifier: 0.773
* RandomForestClassifier: 0.671
* KNeighborsClassifier: 0.700

Как мы видим, дерево решений больше подходят для нашего набора данных, затем k соседей и случайный лес.

В любых возможных сценариях ни один классификатор не работает лучше остальных. На практике всегда реко мендуется сравнить качество, по крайней мере, нескольких разных алгоритмов об учения, чтобы выбрать наилучшую модель для отдельно взятой задачи; алгоритмы могут отличаться по числу признаков либо образцов, уровню шума в наборе данных и по тому, являются классы линейно разделимыми или нет.