Федеральное государственное бюджетное образовательное учреждение высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**Департамент анализа данных и машинного обучения**

Пояснительная записка к курсовой работе

по дисциплине “Машинное обучение”

на тему:

«Машинное обучение в задачах медицинской диагностики»

Выполнил(а):

студент(ка) группы ПМ22-2 факультета информационных технологий и анализа больших данных

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Апанович Д.С.

Научный руководитель:

доцент, к.э.н. Остроухова Н.Г.

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

2024 г

**Введение**

Цель: Изучить теорию и практику применения методов машинного обучения в задачах медицинской диагностики

Задача: Построить несколько моделей для определения пациентов, болеющих раком лёгких, сравнить модели между собой по качеству, эффективности, скорости обучения и т.д. и выбрать среди них наилучшую

**Теоретическая часть**

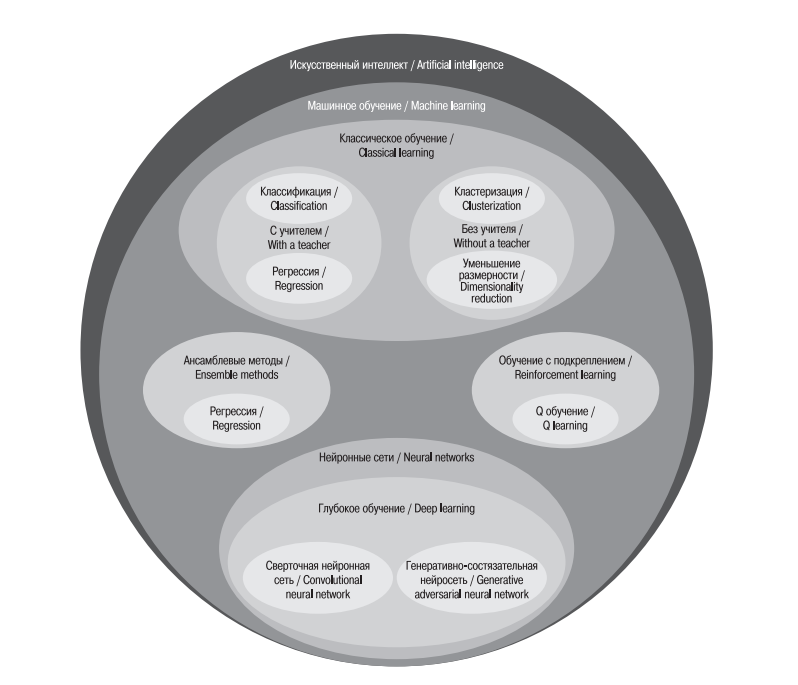
Машинное обучение является относительно новым и перспективным направлением в задачах медицинской диагностики, предоставляя врачам и исследователям мощный инструмент для анализа данных и принятия важных медицинских решений. С постоянным увеличением объема медицинских данных, включая изображения, биомаркеры и клинические записи, методы машинного обучения становятся неотъемлемой частью современной медицины.

Применение алгоритмов машинного обучения в медицинской диагностике позволяет автоматизировать процессы анализа данных, выявлять скрытые закономерности и предсказывать заболевания на ранних стадиях развития. Это способствует более точным диагнозам, оптимизации лечения и улучшению результатов пациентов, а также снижению нагрузки на врачей на различных этапах диагностики и лечения.

Задачи машинного обучения подразделяются на задачи обучения с учителем и без учителя. Модели обучения с учителем подразделяются на задачи регрессии и классификации. В медицинской диагностике наиболее актуальны задачи классификации, однако иногда необходимо находить не качественное, а числовое решение задачи (например, расчёт дозировки лекарства), тогда актуальны задачи регрессии.

В медицине используются как классические модели машинного обучение, так и модели, основанные на глубоком обучении (например, распознавание патологий по снимкам). Перед тем как приступить к построению собственной модели по диагностированию рака лёгких я изучил какие модели машинного обучения использовались другими авторами. Были рассмотрены модели прогнозирования исхода болезни при сердечно-сосудистых нарушениях, предсказания смертельных случаев при сердечной недостаточности, оценки риска рецидива рака.

В основном использовались методы обучения с учителем, такие как: Логистическая регрессия, наивный байес, случайный лес. В более сложных случаях прогнозирование диагноза является задачей многоклассовой классификации.



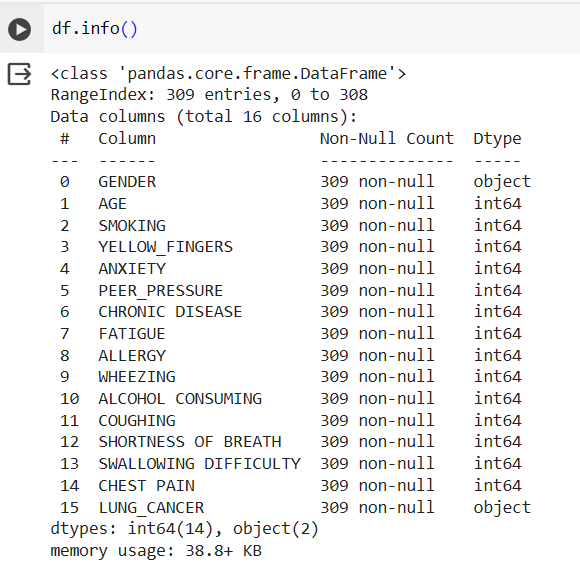
**Практическая часть**

В целях исследования был выбран датасет, содержащий информацию о пациентах с подозрением на рак лёгких, он содержит следующие признаки:

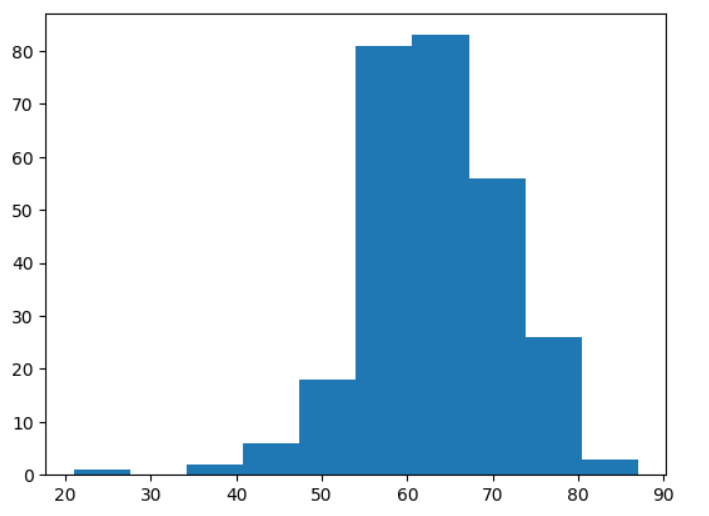
1. Gender: M(male), F(female)
2. Age: Age of the patient
3. Smoking: YES=2 , NO=1.
4. Yellow fingers: YES=2 , NO=1.
5. Anxiety: YES=2 , NO=1.
6. Peer\_pressure: YES=2 , NO=1. (Давление со стороны окружения)
7. Chronic Disease: YES=2 , NO=1. (Наличие хронических заболеваний)
8. Fatigue: YES=2 , NO=1. (Утомляемость)
9. Allergy: YES=2 , NO=1.
10. Wheezing: YES=2 , NO=1. (Хрипота)
11. Alcohol: YES=2 , NO=1.
12. Coughing: YES=2 , NO=1.
13. Shortness of Breath: YES=2 , NO=1.
14. Swallowing Difficulty: YES=2 , NO=1.
15. Chest pain: YES=2 , NO=1.
16. Lung Cancer: YES , NO.

Целевым признаком является последний — наличие или отсутствие рака.

Выведем основную статистическую информацию:

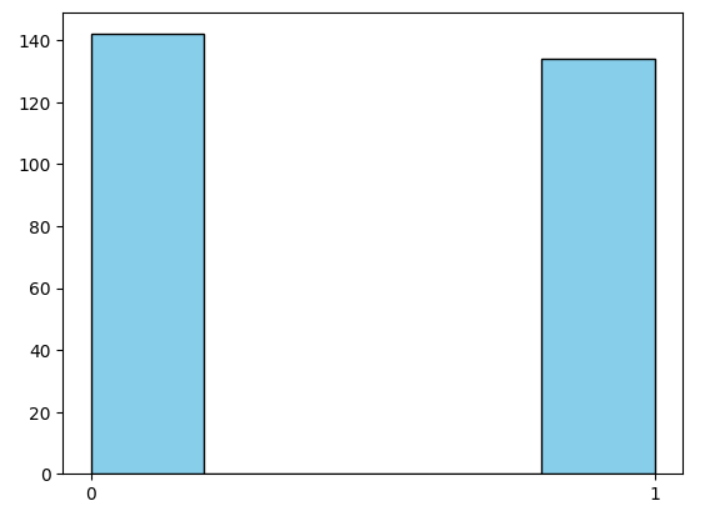
Видим, что данные являются чистыми, нет пропущенных значений, выбросов быть не может, т. к. все признаки являются качественными.

Однако в датасете есть 33 повторяющиеся строки, их необходимо удалить.

Построим гистограмму распределения по возрасту:

Видно, что большинство пациентов — лица пожилого возраста. Это логично, возраст является потенциальным фактором риска, а значит на диагностику чаще всего попадают пожилые пациенты.

Также построим гистограмму распределения по полу:



Примерно равное количество мужчин и женщин.

Визуализация остальных признаков представлена в блокноте.

Построим коррелограмму. Все признаки слабо коррелируют между собой, порог отсечения ни один признак не превышает, поэтому оставляю данные в исходном виде.

Разделим данные. Так как данные не имеют определённой структуры и порядок их следования не важен, применим случайное разделение, датасет небольшой, поэтому выборка поделена в соотношении 85/15

Краткое описание моделей, которые использовались:

*Логистическая регрессия* – основана на логистической функции, которая возвращает вероятность принадлежности к определённому классу.

*Метод k-ближайших соседей* – алгоритм ищет k (задаётся в качестве гиперпараметра) ближайших соседей заданного объекта по метрике расстояния и присваивает этому объекту класс, который преобладает среди соседей

*Гауссовский наивный байес* – в основе работы лежит теорема Байеса, которая позволяет вычислять вероятность принадлежности объекта к определенному классу, основываясь на вероятностях соответствующих признаков (при этом признаки должны быть независимы между собой и нормально распределены)

*Дерево решений* – в этом методе основной идей является разбиение пространства признаков на подмножества, пока не будет достигнут критерий остановки (минимизация ошибки)

*Метод опорных векторов* – целью этого метода является нахождение разделяющей гиперплоскости, которая разделяет данные разных классов и при этом содержит наименьшее количество ошибок. Эта гиперплоскость определяется таким образом, чтобы расстояние до ближайших точек каждого класса (опорных векторов) было максимальным.

*Перцептрон* – работа перцептрона основана на принятии решения о принадлежности объекта к одному из классов на основе взвешенной суммы его входных признаков. Перцептрон принимает входные признаки объекта, умножает их на соответствующие веса и складывает их. Затем полученная сумма проходит через функцию активации, которая определяет, к какому классу отнести объект.

*Случайный лес* – это ансамбль моделей деревьев решений, основная идея которого заключается в том, что он строит не одно дерево решений, а множество деревьев. Каждое дерево строится независимо от других на основе случайной подвыборки данных и случайного подмножества признаков. Затем для классификации или регрессии используется голосование или усреднение результатов всех деревьев.

Я построил модель по каждому из этих методов и получил по основным метрикам следующие результаты:

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Accuracy** | **Precision** | **Recall** |
| Логистическая регрессия | 0,88 | 0,868 | 1 |
| Метод k-ближайших соседей | 0,785 | 0,8 | 0,969 |
| Гауссовский наивный байес | 0,88 | 0,888 | 0,969 |
| Дерево решений | 0,857 | 0,909 | 0,909 |
| Метод опорных векторов | 0,785 | 0,785 | 1 |
| Перцептрон | 0,69 | 0,95 | 0,636 |
| Случайный лес | 0,904 | 0,914 | 0,969 |

Для медицинской диагностики наиболее важна метрика recall, т.к. цена ошибки при ложноотрицательном диагнозе гораздо больше, чем при ложноположительном. Кроме того, по каждой модели были построены кривые обучения.

Для дальнейшего исследования я выбрал модели логистической регрессии и случайного леса, как наиболее перспективные.

Для выбранных моделей была проведена кросс-валидация с 4 фолдами, обе модели показали примерно равный результат по метрике recall.

Далее я провёл поиск наилучших гиперпараметров по сетке (Grid Search) и получил следующие результаты

**Логистическая регрессия**

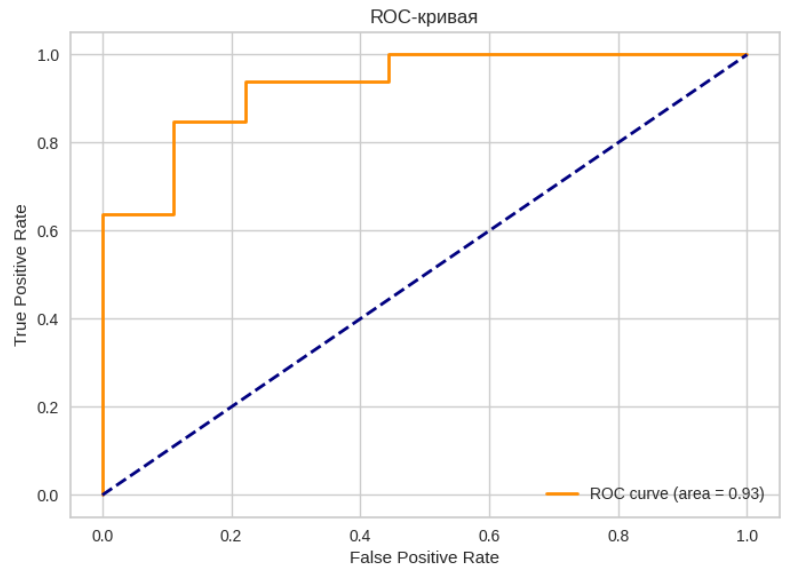
|  |  |
| --- | --- |
| **Параметр** | **Значение** |
| penalty | L1 |
| C | C=3.359 |
| solver | saga |
| max\_iter | 1000 |

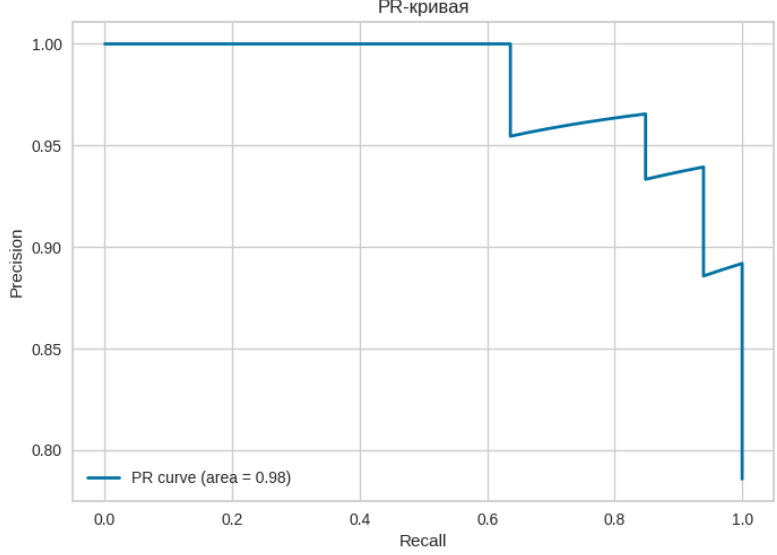
**Случайный лес**

|  |  |
| --- | --- |
| **Параметр** | **Значение** |
| n\_etimators | 100 |
| max\_features | sqrt |
| max\_depth | 6 |
| max\_leaf\_nodes | 9 |

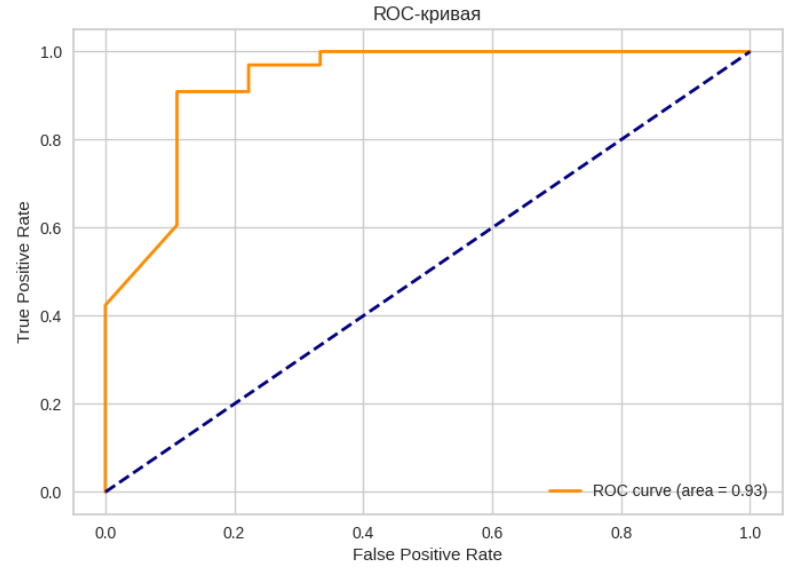
После этого были построены roc и pr кривые для обеих моделей для оценки производительности

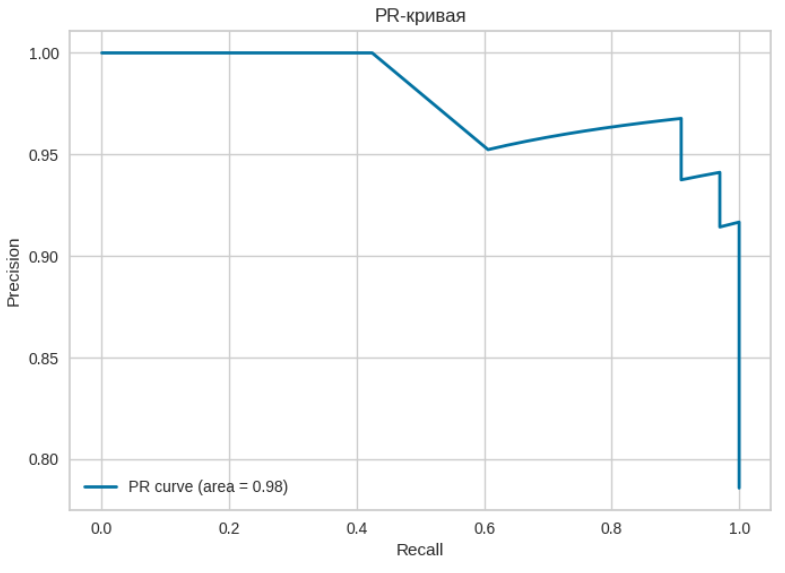
**Логистическая регрессия**

****

****

**Случайный лес**

****

****

**Вывод**

В ходе работы были изучены и применены на практике 7 моделей классического машинного обучения, из них эффективными оказались две: логистическая регрессия и случайный лес. Несмотря на примерно равные показатели по метрикам, по-моему мнению, предпочтение стоит отдать модели логистической регрессии, потому что она проще и будет работать быстрее на больших объёмах данных. В целом, применение моделей машинного обучения в медицинской диагностике является перспективным, они позволят снизить нагрузку на врачей, а также минимизировать человеческий фактор. Однако, для того чтобы их можно было эффективно применять, необходимо получать на вход чистые и достоверные данные, сбор которых также является проблемой, решить которую довольно трудно.

### Список литературы

Основная литература

* [М.В. Коротеев. Об основных задачах дескриптивного анализа данных.](https://drive.google.com/file/d/1JoHxg3dfc53bSZtsNfPPB5Es_C4Q9hJ0/view?usp=sharing)
* [М.В. Коротеев. Учебное пособие по дисциплине “Анализ данных и машинное обучение” - 2018.](https://portal.fa.ru/Files/Data/85bca4b4-cd40-4dae-bb62-dbd6c04fa0a0/um_tehnanalizadannihimashobuchenie_18.pdf)

Дополнительная литература

* A. Geron. Hand on Machine Learning with scikit-learn and Tensorflow - 2017 (564p)
* C. Albon. Machine learning with Python Handbook - 2018 (427p)
* L.P. Coelho, W. Richert. Building machine learning systems with Python - 2015 (326p)
* J. Grus. Data science from scratch - 2015 (330p)
* W. McKiney. Pandas: powerful Python data analysis toolkit - 2016 (1971p)

Видео-ресурсы

* [Классический курс по машинному обучению](https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN)
* [Конспекты](http://cs229.stanford.edu/syllabus.html)
* [Материалы на гитхабе](https://github.com/vkosuri/CourseraMachineLearning)
* [Pandas tutorial](https://www.youtube.com/playlist?list=PLeo1K3hjS3uuASpe-1LjfG5f14Bnozjwy)
* Канал [Python programmer](https://www.youtube.com/user/consumerchampion/playlists) - много контента по программированию на питоне
* Серия [Data analysis](https://www.youtube.com/playlist?list=PLzH6n4zXuckpfMu_4Ff8E7Z1behQks5ba) от Computerphile
* [ИИ и машинное обучение в медицине / Хабр (habr.com)](https://habr.com/ru/companies/cloud4y/articles/506288/)
* [Применение машинного обучения в диагностике туберкулеза: систематический обзор литературы (cyberleninka.ru)](https://cyberleninka.ru/article/n/primenenie-mashinnogo-obucheniya-v-diagnostike-tuberkuleza-sistematicheskiy-obzor-literatury/viewer)
* [Machine learning in analysis of biomedical and socio-economic data (spbstu.ru)](https://elib.spbstu.ru/dl/2/id20-61.pdf/download/id20-61.pdf?ysclid=lvqh35oqng59812002)