4 части работы

1. Медицина
2. Анализ данных
3. Нейросетевые методы обработки
4. Написание интерфейса
5. Добавить и изменить по статье
6. Теория про примененные методы и общая ИНС
7. .
8. .

**Введение**

Актуальность:

- Развитие телемедицины

- Болезнь (рак легких) не имеет лекарств и ее невозможно предсказать

- Нежелание регулярно проходить обследования у врачей, и далеко не у каждого имеется такая возможность

- В будущем можно будет внедрить систему предупреждения о возможном риске заболевания из базы пациентов (предупреждение тех, кто не заботится о себе)

- Скрининговой (профилактической) программы обследования лиц, входящих в группу риска по раку лёгкого, в нашей стране пока нет.

Рак – на него приходится около 25% всех смертей в России, причем наиболее распространенными видами являются рак легких и рак желудка.

Целью работы является анализ набора данных об известных случаях заболевания, выявление наиболее информативных симптомов, разработка интерфейса по выявлению риска заболевания у отдельного пользователя

Результат работы данного приложения не будет являться диагнозом специалиста, но будет сигнализировать о нахождении в группе риска и рекомендовать записаться к врачу. В перспективе приложение можно будет дополнять другими заболеваниями для общего контроля здоровья, к примеру, инсульт, на который приходятся рекордные 40% смертей.

**Часть 1 Заболевание**

Первым делом необходимо установить основные симптомы и специфику заболевания, чтобы понять, какие данные нужно собирать с новых пользователей

**Распространенность**

Среди злокачественных новообразований рак лёгкого занимает лидирующие позиции в мире как по своей встречаемости, так и по смертности. Вот уже более 30 лет он остаётся главным "онкологическим киллером". Мужчины болеют в 6 раз чаще, чем женщины. За последние 20 лет общая заболеваемость выросла более чем в 2 раза, при этом среди мужчин на 51%, а среди женщин на 75%.

Средний возраст, в котором диагностируется рак лёгкого, достигает 71 года. Реже всего заболевание обнаруживается у людей младше 20 лет.

**Этиология**

* Курение
* Профессиональные факторы
* Хронические заболевания легких
* Возраст свыше 45 лет

**Симптомы и Этиология**

У рака лёгкого нет специфических симптомов, а в 15% случаев болезнь протекает бессимптомно.

Первые признаки рака лёгкого

* Хронический кашель
* Боль в груди
* Хрипота
* Одышка
* Слабость в конечностях
* Появление желтизны

Конечно, мы уже говорим о болезни непосредственно. В реальности, в группе риска данного заболевания находится каждый человек, потому что на ранних стадиях болезнь протекает бессимптомно, и она может возникнуть без весомых на то причин.

Главной задачей данного раздела является выбор ведущих факторов, влияющих на развитие онкологического заболевания. Мы не стремимся определять нахождение в группе риска по 100 или более признакам, это было бы бессмысленно, поскольку пользователь предпочтет сходить к врачу, чем тратить много времени на заполнение полей опроса. Но и оставлять 2-3 фактора нельзя, доверие пользователя и польза от такого анализа минимальны.

**Аллергия**

По исследованиям, аллергия скорее враг рака, чем друг. При повышенном уровне иммуноглобулина Е (IgE) организму свойственно остро реагировать на любые сильные возбудители и выводить их из организма. Как раз такие возбудители могут вызвать метаморфозы в клеточной структуре органов человека и пробудить раковые клетки. В этом несложно проследить логику.

**Курение**

По результатам многочисленных исследований давно выявлено не решающее влияние курение на риск развития рака. Курение скорее всецело состаривает организм и вызывает смертность клеток. Тем более что в условиях большого города (мы не имеем данных о локализации собранных данных) каждый человек получает вредное воздействие на легкие загрязненной атмосферой или пассивным курением, которое ничем не лучше обычных сигарет. Также почти каждый человек хоть раз в жизни пробовал курение, а как известно 1 капля никотина убивает лошадь, а в пересчете на сигареты это в районе 2.5 пачек. Большинство курильщиков выкуривает такое количество за 2-3 дня. Именно поэтому нет разумных причин включать курение в список анализа.

**Возраст**

**Алкоголь**

**Хрипота**

Данный фактор уже относится к симптомам и может служить флажком на ранних-средних стадиях развития заболевания

**Боль в груди**

Воспаление легкого

**Кашель**

Воспаление дыхательных путей, трахеит и тд

=Теоретическая база про ИНС в общем и в задачах классификации=

* Каскадные сети как способ избежания дисбаланса

**Часть 2 Анализ данных (EDA)**

Проведем первичный анализ набора данных об исследуемом заболевании

**Описание набора данных Survey lung cancer**

<https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer>

Набор содержит данные по 309 респондентам, каждый из которых обследовался на рак легких и получил впоследствии диагноз.

Данные собираются с сайта онлайн-системы прогнозирования рака легких и получают обратную связь от пользователя. Этот сайт был реализован в течение августа 2013 года людьми, посетившими этот сайт.

Набор признаков:

- GENDER (пол): M1 (Муж), F2 (Жен)

- AGE (Возраст): Непрерывная величина возраста пациента

- SMOKING (Курение): Yes (Да) =2, No (Нет) =1

- YELLOW\_FINGERS (Желтые пальцы): Yes (Да) =2, No (Нет) =1

- ANXIETY (Тревожность): Yes (Да) =2, No (Нет) =1

- PEER\_PRESSURE: Yes (Да) =2, No (Нет) =1

- CHRONIC DISEASE (Хронические заболевания): Yes (Да) =2, No (Нет) =1

- FATIGUE (Усталость): Yes (Да) =2, No (Нет) =1

- ALLERGY (Аллергия): Yes (Да) =2, No (Нет) =1

- WHEEZING (Хрип): Yes (Да) =2, No (Нет) =1

- ALCOHOL CONSUMING (Употребление алкоголя): Yes (Да) =2, No (Нет) =1

- COUGHING (Кашель): Yes (Да) =2, No (Нет) =1

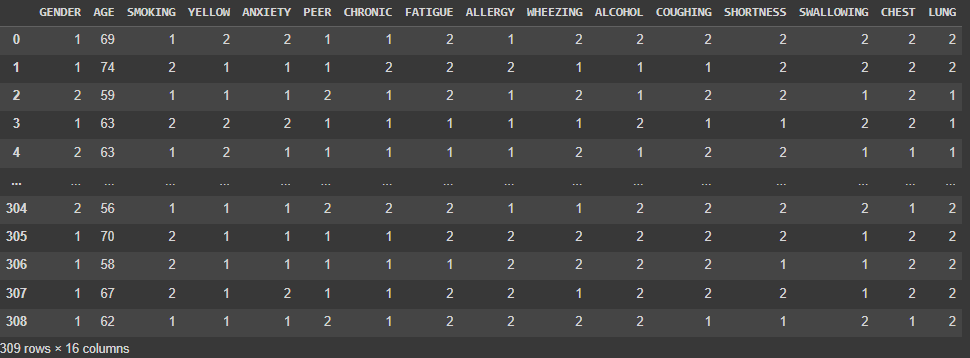
- SHORTNESS OF BREATH (Одышка): Yes (Да) =2, No (Нет) =1

- SWALLOWING DIFFICULTY (Затрудненное глотание): Yes (Да) =2, No (Нет) =1

- CHEST PAIN (Боль в груди): Yes (Да) =2, No (Нет) =1

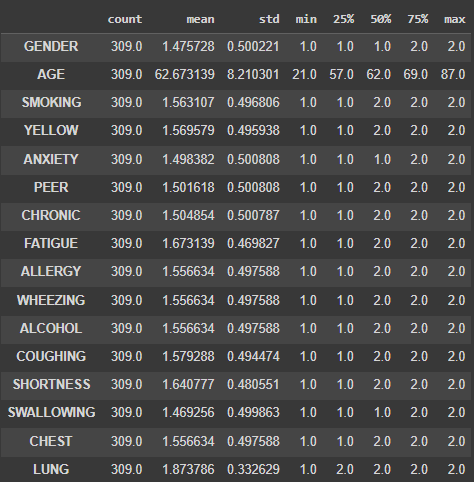
Зависимая переменная:

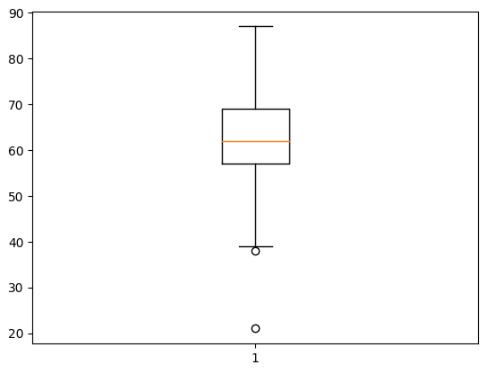
- LUNG\_CANCER (Рак легких): Yes (Да) =2, No (Нет) =1



**Разведочный анализ данных (EDA)**

Все переменные, кроме возраста, имеют группировочный анализ по двум признакам (да или нет). Имеет место предположение о том, что для хорошего обучения (впоследствии) модели-классификатора будет идеальным равномерное распределение группировочных переменных и нормальное – непрерывных.





Хорошее по возрасту (близ нормальное)

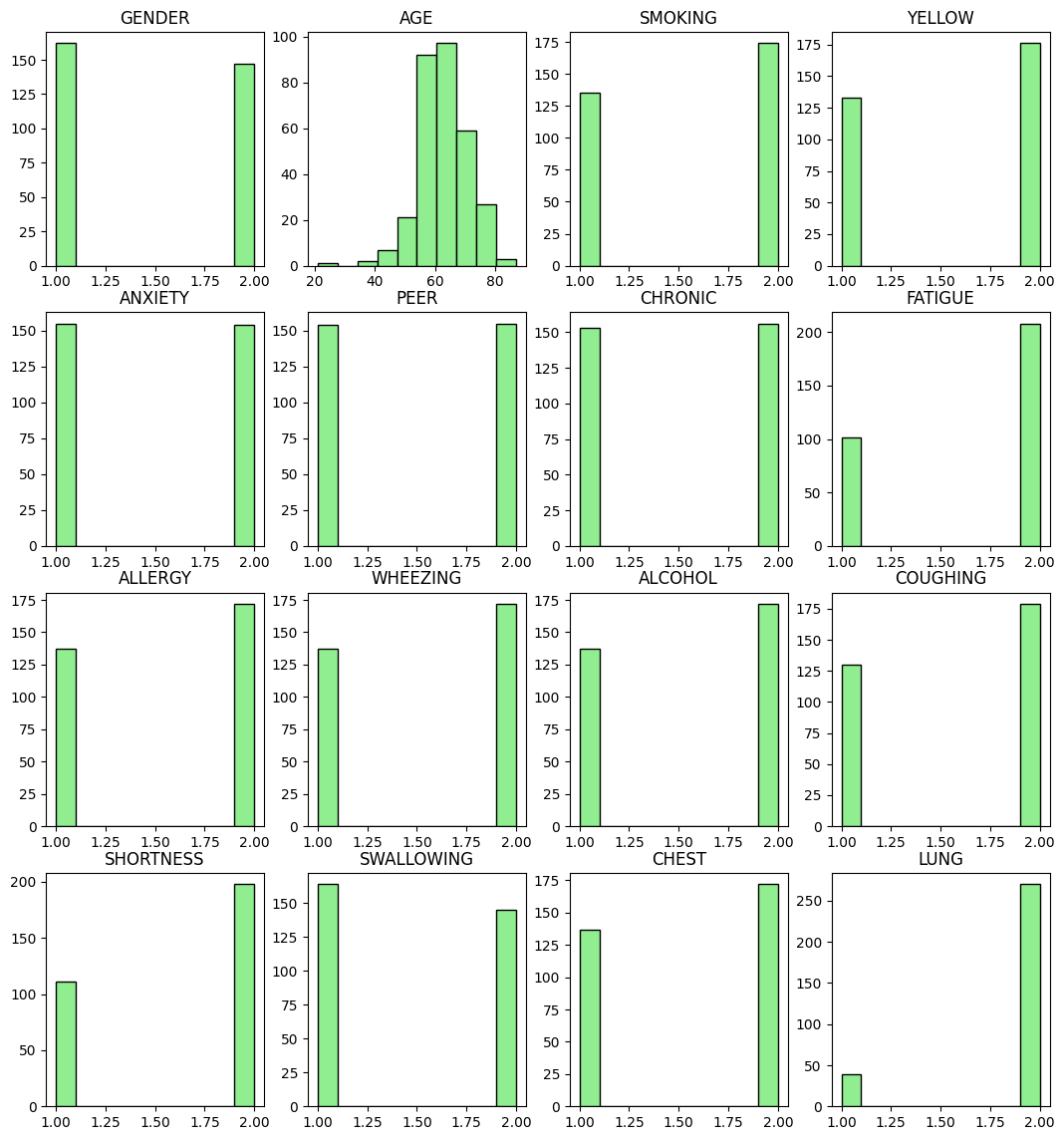


Рис. i. Распределение данных

По распределениям данных можно лишь сделать вывод о несбалансированности наших данных, то есть зависимая переменная диагноза распределена неравномерно, что может привести к перевесу обучения в сторону класса с преимуществом по числу пациентов.

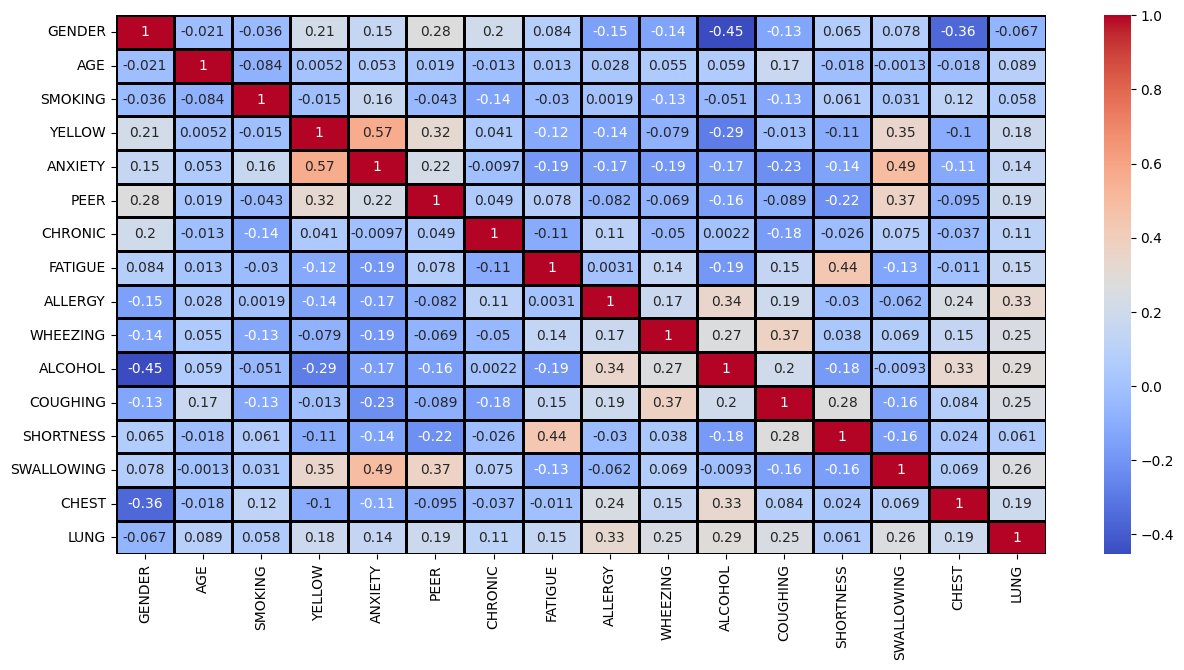
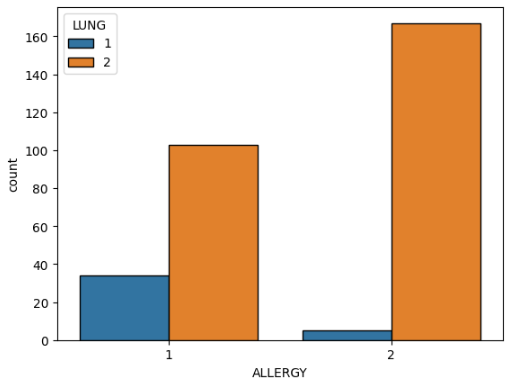


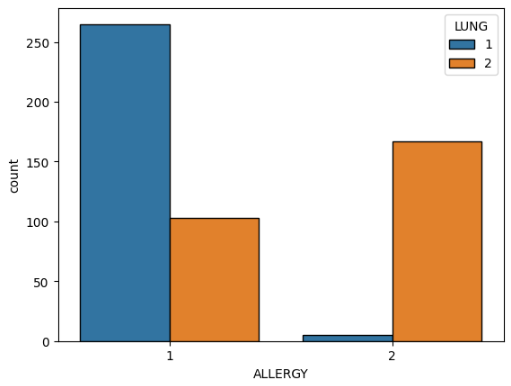
Рис. i. Heatmap\_lung

Мы знаем, что из приведенного выше графика корреляции

1. Курение имеет низкую корреляцию со всеми переменными, влияющими на рак легких, а также не коррелирует с раком легких.
2. Желтые пальцы и беспокойство имеют высокую корреляцию, которая составляет 57%, и оба имеют достойную корреляцию с раком легких.
3. Затрудненное глотание и беспокойство имеют высокую корреляцию с 49% и каким-то образом хорошо коррелируют с раком легких.
4. Одышка и усталость имеют высокую корреляцию с 44% и имеют достойную корреляцию с раком легких.
5. Аллергия и употребление алкоголя имеют высокую корреляцию с 34% и самую высокую корреляцию с раком легких, что указывает на то, что это важный фактор рака легких.
6. Употребление алкоголя и свистящее дыхание также имеют высокую корреляцию с 37% и высокую корреляцию с раком легких, что также указывает на фактор рака легких.
7. Боль в груди и употребление алкоголя также имеют высокую корреляцию с 33%, а также сильно коррелируют с раком легких, что также может указывать на фактор рака легких.



Видим что отсутствие аллергии так же не говорит об отсутствии рака, можно сказать что сильная коррелированность аллергии и диагноза обусловлена тем, что у нас преимущество по



После применения SMOTENC видим, что отсутствие аллергии теперь с большей вероятностью говорит об отсутствии рака. Посмотрим корреляционную матрицу теперь. По идее, она должна измениться мало (тк сэмплирование предполагает сохранение имеющихся закономерностей, если они нарушены, следует изменить метод передискретизации)

**SMOTENC**

Для сэмплирования набора данных был выбран метод SMOTE, но он создает новые экземпляры с помощью интерполяции между известными объектами, что неприменимо при категориальных данных, так как 1 и 2 являются метками классов. У метода SMOTE существует две модификации: SMOTEN и SMOTENC. Первый метод создан только для категориальных данных, но наш набор содержит также одну непрерывную переменную, а именно возраст, значит нам следует использовать метод SMOTENC, который работает смешанно, следует лишь пометить, какие данные, подаваемые на вход, являются категориальными. Суть работы данного метода почти не отличается, работает так же по методу k-ближайших соседей.

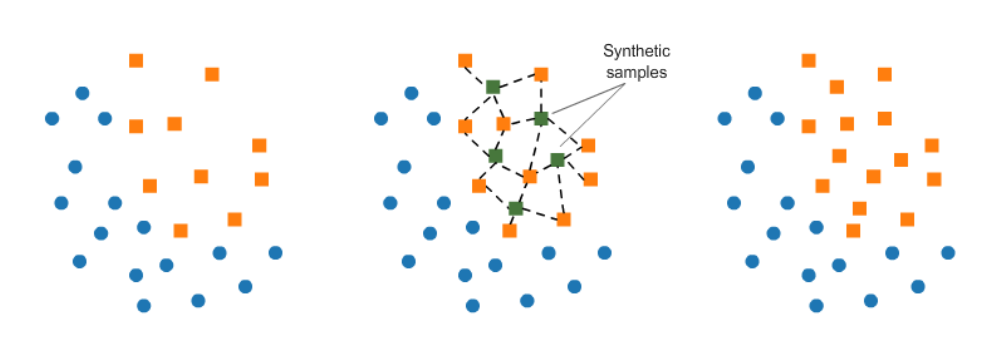


Рис. i. Графическое пояснение интерполяции при сэмплировании

Алгоритм SMOTE:

1. Выберите образец класса меньшинства из исходного набора данных.
2. Найдите k ближайших соседей по классу меньшинства в пространстве объектов.
3. Случайным образом выберите одного из k ближайших соседей.
4. Создайте новую синтетическую выборку путем интерполяции между выбранной выборкой класса меньшинства и случайно выбранным соседом.
5. Повторяйте шаги 1-4, пока не будет сгенерировано желаемое количество синтетических выборок.

SMOTENC немного изменяет способ создания нового сэмпла, выполняя что-то определенное для категориальных функций. Фактически, категории новой сгенерированной выборки определяются путем выбора наиболее часто встречающейся категории ближайших соседей, присутствующих во время генерации, и при вычислении евклидова расстояния при поиске k-ближайших-соседей алгоритм просто добавляет квадрат медианы стандартных отклонений каждого столбца, представляющего числовые значения.

**Выводы по анализу**

**Применение методов ИАД**

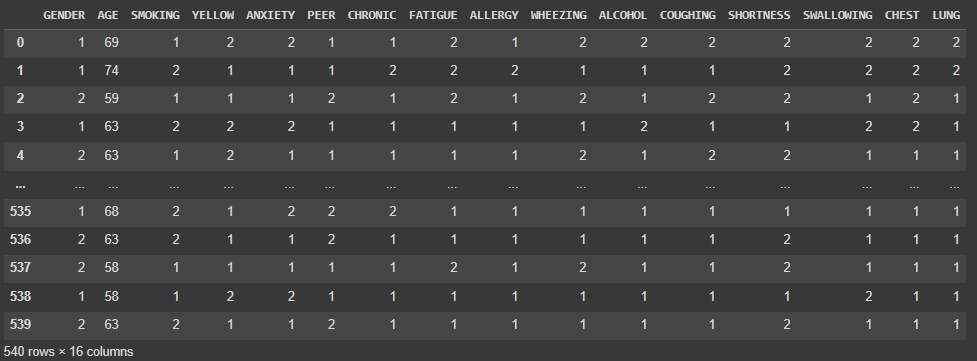
Модель-классификатор по выявлению риска

Модель логистической регрессии

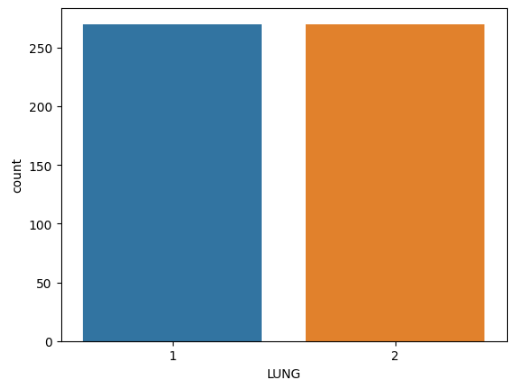
Моделирование возможно только на сбалансированных данных. Потому что, если данные о дисбалансе произошли, модель будет иметь смещение, потому что данные, как правило, показывают больше данных «LUNG\_CANCER», которые нужно обучить.

Это означает, что он с большей вероятностью будет предсказывать будущие данные в метке большинства.

Использую метод SMOTENC по передискретизации



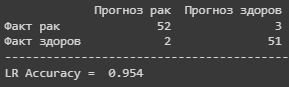
Итого получаем 540 пациентов, среди которых 270 больны и 270 здоровы



Разделим наши данные на обучающие и тестовые данные в соотношении 4:1 (80% : 20%) 432 / 108

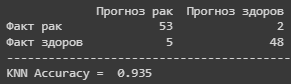
**Logistic Regression (Логистическая регрессия)** далее LR

==Теория==



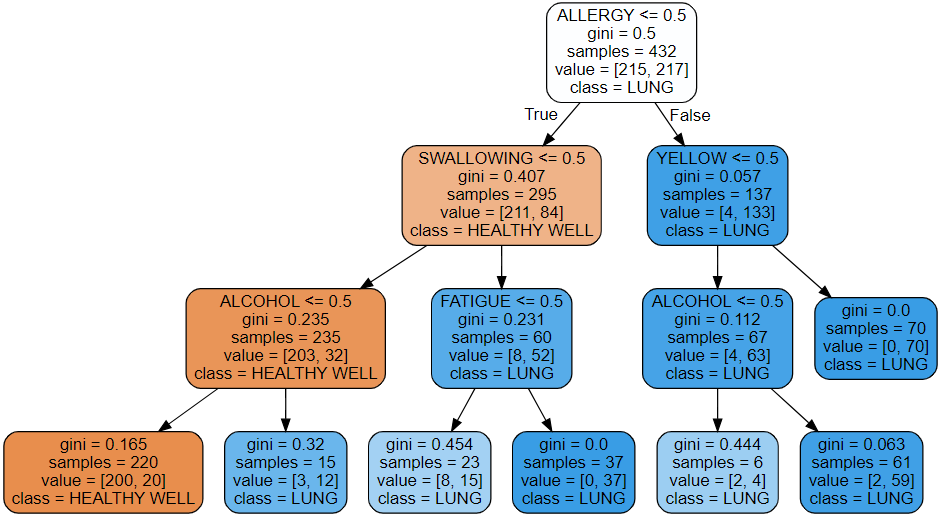
**K-Nearest Neighbors (K-ближайших соседей)** далее KNN

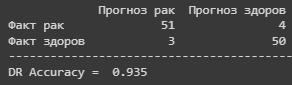
==Теория==



**Decision Tree (Дерево решений)** далее DT

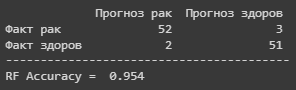
==Теория==





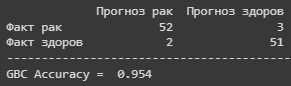
**Random Forest (Случайный лес)** далее RF

==Теория==



**Gradient Boosting Classifier (Градиентный бустинг)** далее GBC

==Теория==



**Анализ результатов**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Метод | LR | KNN | DT | RF | **GBC** |
| Точность | 0.954 | 0.935 | 0.935 | 0.954 | 0.954 |

В данной задаче перед нами стоит цель добиться создания наиболее точной модели предсказания рака легких. Следовательно, при создании ИНС мы оставляем в данных все 15 предсказательных атрибутов.

**Выводы по анализу данных**

**Часть 3 Создание модели ИНС**

**Выбор архитектуры**

Мы имеем достаточно простую информацию, обрабатывать которую сможет ИНС прямого распространения с несколькими скрытыми слоями.

Входной слой обязан принимать на себя 15 признаков, значит будет содержать 15 нейронов

Выходной слой должен содержать столько нейронов, сколько классов мы имеем, т.е. 2

Попробуем следующие вариации тестирования лучшей модели:

- Что делать с архитектурами, какие сравнивать

* По архитектуре (кол-во слоев и нейронов, ф-ии активации выходного и скрытых слоев)
* После по параметрам обучения (эпохи,
* Задание двойного приоритета на решающие факторы предсказания аллергия, алкоголь

Поскольку перебрать все комбинации различных вариаций ИНС не представляется возможным, предлагается соблюдать движение по списку, т.е. Находить лучший по архитектуре, далее подгонять по параметрам обучения, после чего задавать приоритет при обучении.

Наша цель - достичь максимальной точности на тестовой выборке, чтобы потеря была минимальной.

**Обучение модели**

Параметры обучения:

Для отслеживания переобучения разумно выделить из обучающих данных 10% на валидационные

В качестве начального значения эпох рассмотрим 1000, дальше будем смотреть

функция активации выходного слоя: softmax

функция ошибки: categorical\_crossentropy

алгоритм обучения: sgd

метрика качества: accuracy

Данные необходимо представить в виде, удобном для обработки нейросетью

**Моделирование**

Создаем модель Sequential

**Анализ результатов**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Количество скрытых слоев | Количество нейронов в первом скрытом слое | Количество нейронов во втором скрытом слое | Количество нейронов в третьем скрытом слое | Значение метрики качества классификации |
| 0 | - | - | - | 0.9352 model\_1 |
| 1 | 10 | - | - | 0.9352 model\_10 |
| 20 | 0.9537 model\_20 |
| 2 | 20 |  |  |  |
|  |  |
|  |  |
|  |  |
| 3 | 20 |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Приведем результат работы лучшей архитектруры \*()

Тестовый больной

Тестовый здоровый

Неверное предсказание

Как видим, значения на выходе ИНС близки, возник конфликт и предсказание оказалось неверным.

Вывод:

Лучшая инс по показателю метрики качества accuracy:

Она очень проста, но и наши данные в целом достаточно простые

**Выводы по моделированию ИНС**

**Часть 4 Интерфейс**

Идея данного приложения не имеет смысла создания exe-файла, т.к. приложение носит слишком краткосрчный и информативный характер, пользователь не нуждается в ежедневном и длительном использовании. Для ПК оптимальным будет веб-страница, которую можно легко посетить, провериться и закрыть. В случае смартфона уже можно подумать о настольном приложении, чтобы обеспечить память по введенным данным, установку напоминаний о необходимости ввести новые данные, но это нами не рассматривается. Например, можно добавить ввод веса пациента и когнтролировать его по неделям; Если наблюдается регресс (потеря веса вследствие потери воды в организме), предлагать посетить врача.

Суть:

После открытия пользователь пользователь имеет возможость заполнить 15 пунктов, в результате которых уже обученная нейросеть сможет выдать предполагаемый диагноз пользователю. Кроме того, если приложение не ошиблось, данные о пользователе могут быть собраны и использованы далее для следующего обучения ИНС.

Так же хочется добавить базу пользователей с аккаунтами, чтобы пользователь мог в любой момент указать появление одного из симптомов и сразу узнать новый диагноз

ТЗ на интерфейс:

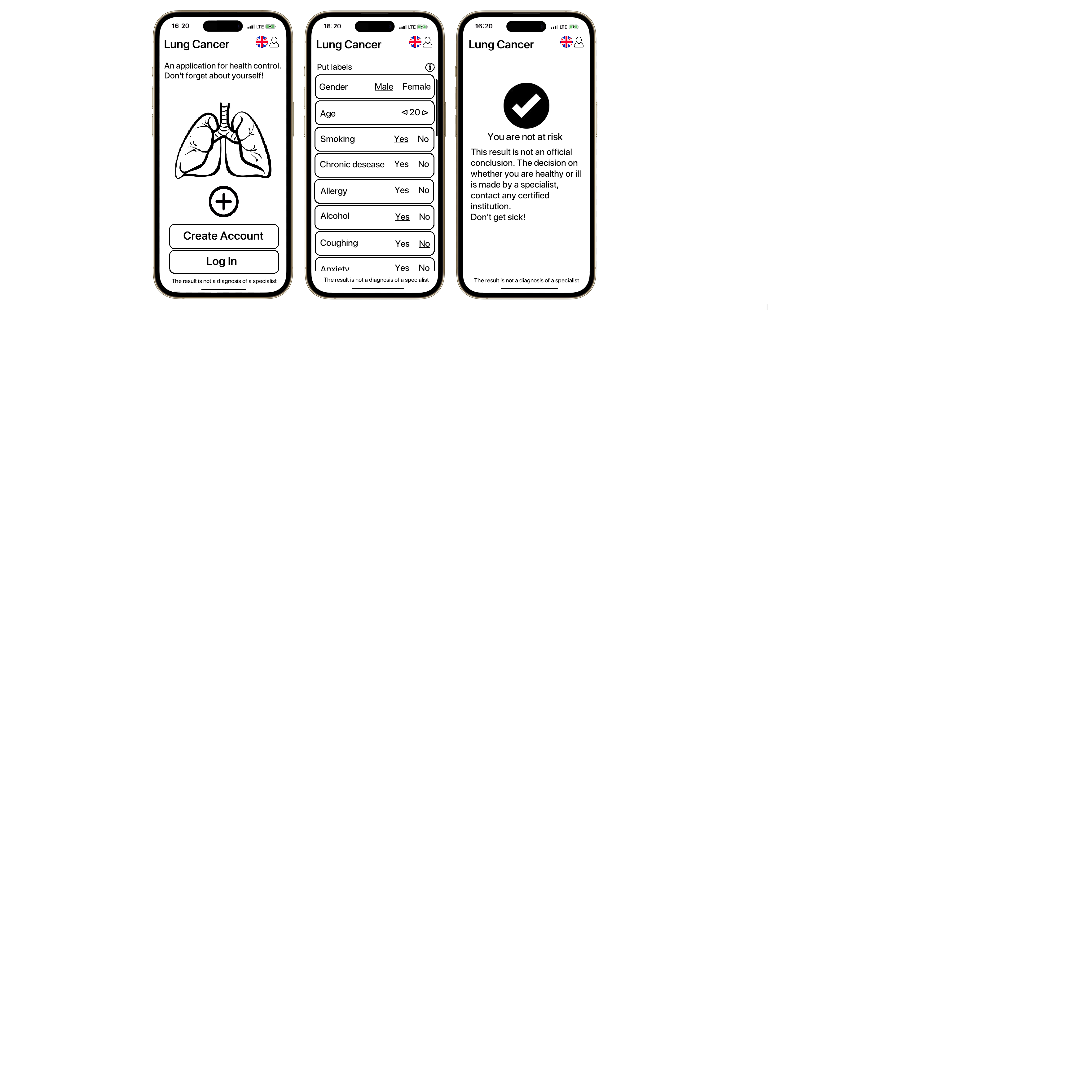
В среду разработки VS code имеет добавлять лишь выбранную наилучшую ИНС, весь первичный анализ EDA будет отягощать программу и не нести в себе полезности.

* левое меню с вкладками: проверка, заболевание (краткая информация про заболевание и профилактику), ..., справка
* Центрированный ввод данных (Да/Нет)

В качестве библиотеки для написания веб страницы мною выбрана библиотека Django, которая имеет простое описание и широкие возможности использования

Django — это высокоуровневый Python веб-фреймворк, который позволяет быстро создавать безопасные и поддерживаемые веб-сайты. Созданный опытными разработчиками, Django берёт на себя большую часть хлопот веб-разработки, поэтому вы можете сосредоточиться на написании своего веб-приложения без необходимости изобретать велосипед.

Ниже представлен эталон интерфейса на мобильный телефон

****

Возможный мобильный удобный интерфейс