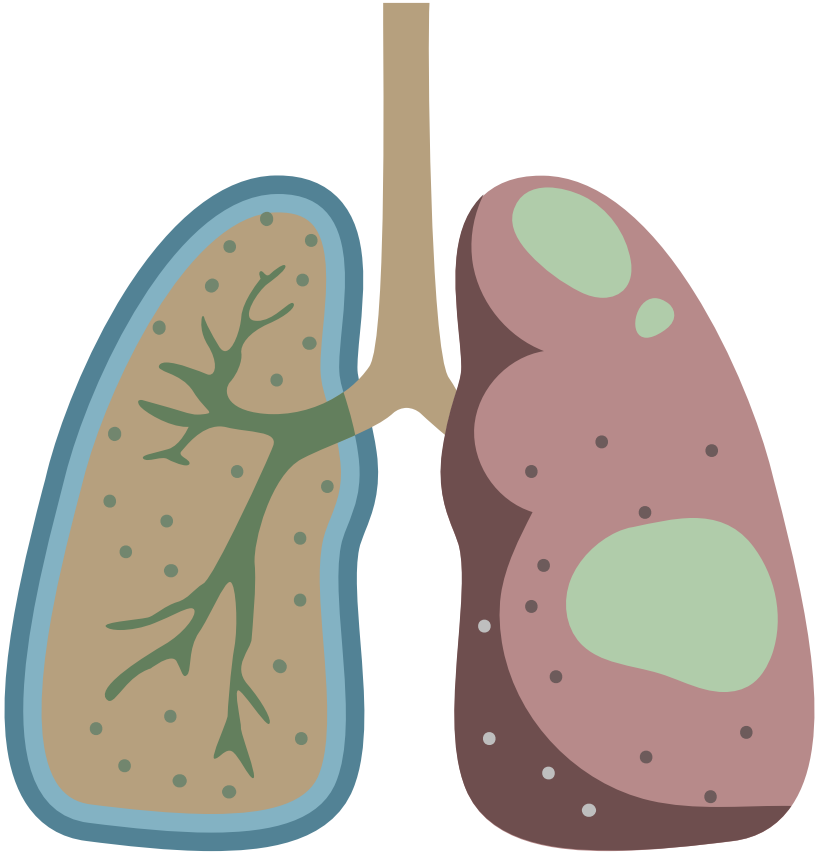
Respiratory Risk Patterns (RRP)

Определение паттернов возникновения заболеваний органов дыхания
и обучение модели предсказания риска развития заболевания



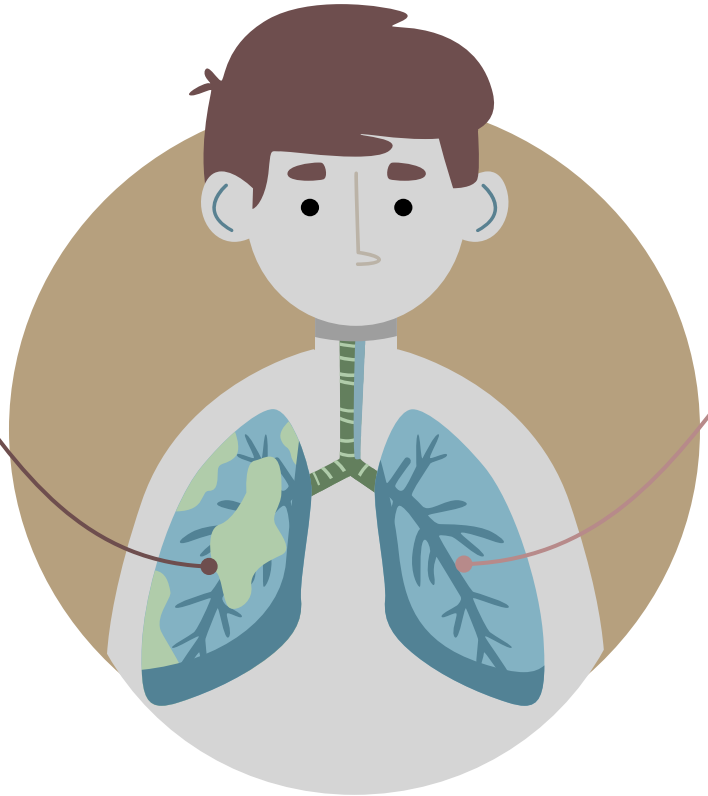
Актуальность проекта

A

Проблема

Респираторные заболевания являются причиной смерти в 12,1% случаях

Рак лёгких часто является причиной смерти из-за поздней диагностики, а не высокой степени опасности

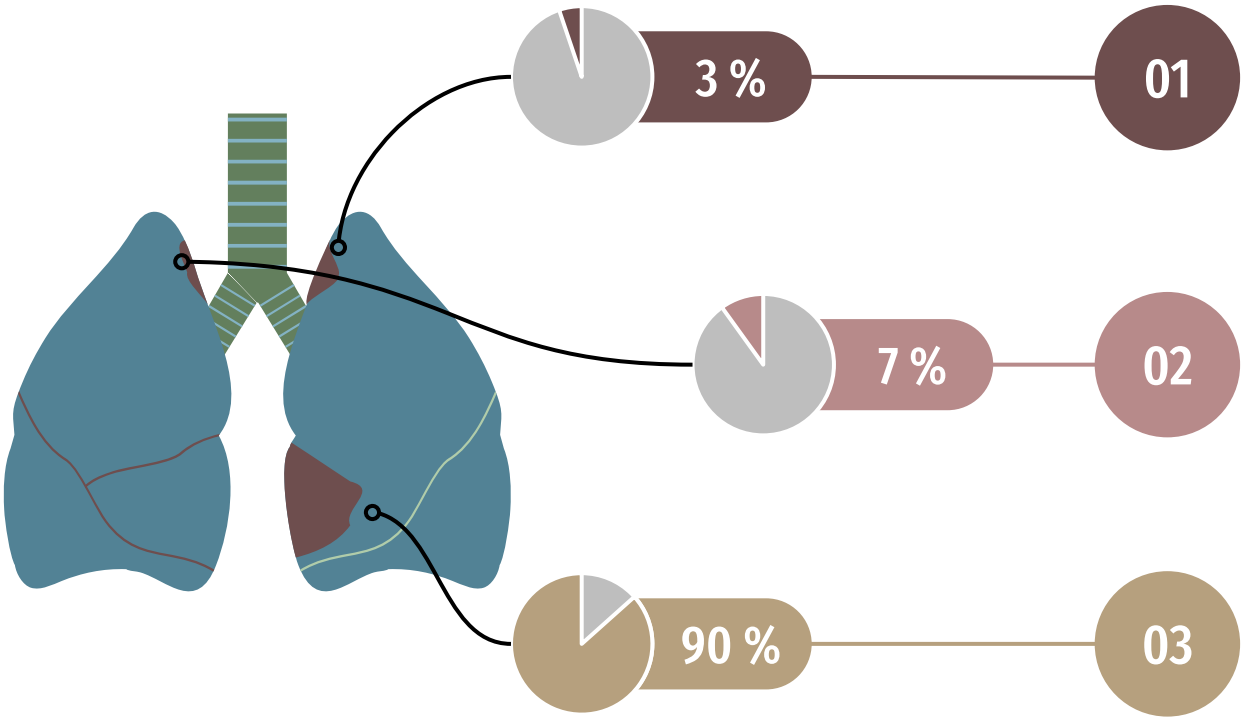


B

Решение

Система быстрой диагностики, предсказывающая риск развития неинфекционного респираторного заболевания в зависимости от образа жизни и особенностей респондента

Сбор данных
(датасеты с сайта kaggle.com)



ХОБЛ

занимает 2-е место среди причин смерти людей по всему миру

Рак лёгких

зачастую является причиной смерти из-за поздней диагностики

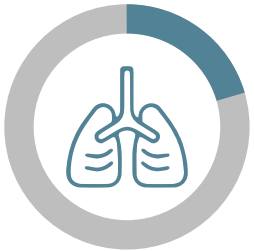
Сводный

датасет жителей США, содержит информацию по ХОБЛ, раку лёгких и бронхиальной астме

Предобработка данных (файл rename.py, вызываемый main.py)

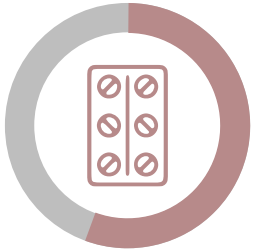
df['feature'].replace()

Стандартизация значений признаков



df.rename()

Стандартизация названий признаков



(df[list]==1).any(axis=1)

Создание единого признака “хронические заболевания”

df['feature'].apply()

Создание переменной “респираторные заболевания”

Интеграция данных
(база данных - combined_dataset.db создана main.py)

gender

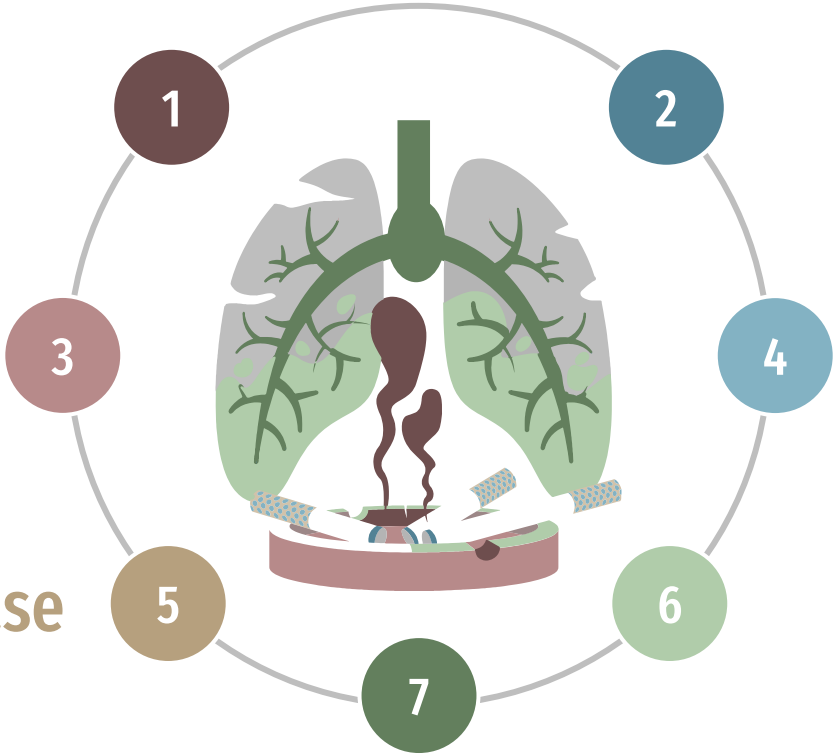
age

smoking

diabetes

chronic_disease

alcohol



pulmonary_diseases

Оценка качества данных (файл statistics_summary.csv создан analisis.py)

	count	mean	std	min	max	mode	unique	missing, %
gender	372147	0.46	0.50	0	1	0	2	0
age	372147	54.53	17.69	18	88	80	67	0
smoking	372147	1.68	0.87	1	3	1	3	5.24
diabetes	372147	0.16	0.37	0	1	0	2	0.28
chonic disease	372147	0.55	0.50	0	1	1	2	0
alcohol	372147	1.51	0.50	1	3	2	3	6.65
pulmonary diseases	372147	0.54	1.16	0	4	0	4	0

Оценка качества данных. Согласованность

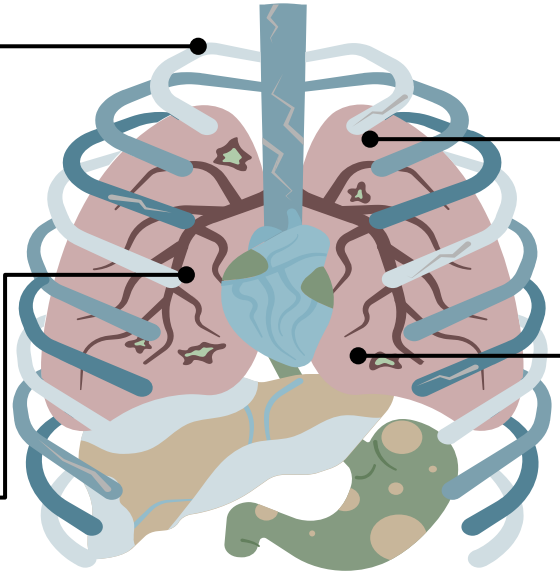
В разных датасетах пол, статус курения, употребление алкоголя, наличие диабета и хронических заболеваний обозначалось по-разному, поэтому данные по этим колонкам подверглись стандартизации, изначально данные были плохо согласованы

Пол

0/1 or 1/2 or M/F

Курение, алкоголь, диабет

0/1 or Y/N or 1/2/3/4/7/9 (добавляется статус “бывший”)



Респираторные заболевания

0/1 or several columns

Хронические заболевания

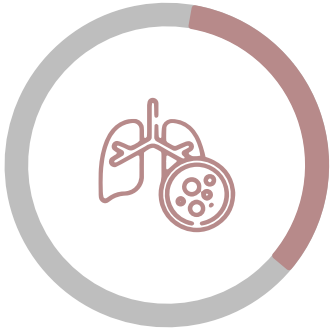
0/1 or several columns

Анализ данных
(результаты в папке results, созданы файлом analisis.py)

6

Распределения

всех признаков, по которым идёт анализ



5

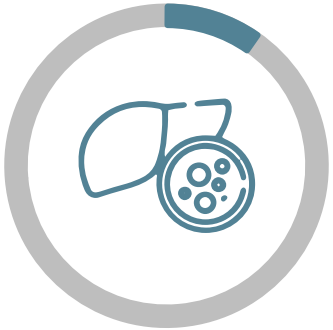
Зависимости

наличия респираторного заболевания от признаков

1

Частота

встречаемости каждого заболевания



2

Корреляция

всех признаков, в том числе наличия заболевания

Найденные паттерны (графики представлены в папке results)

Курение, алкоголь

Повышают вероятность развития респираторного заболевания

Диабет

Повышает риск заболеть на 27%

Пол, возраст

У мужчин на 17% больше риск заболеть. Общий риск повышается с возрастом

Обучение модели (файл RF.py)

RandomForestClassifier
test_size=0.3, random_state=42, stratify=y

Целевая переменная - pulmonary_diseases,
остальные переменные - признаки

Точность предсказания модели - 83,3%

Ключевые факторы: хронические заболевания,
возраст, курение (общая значимость - 90%)

Заключение

Сбор данных

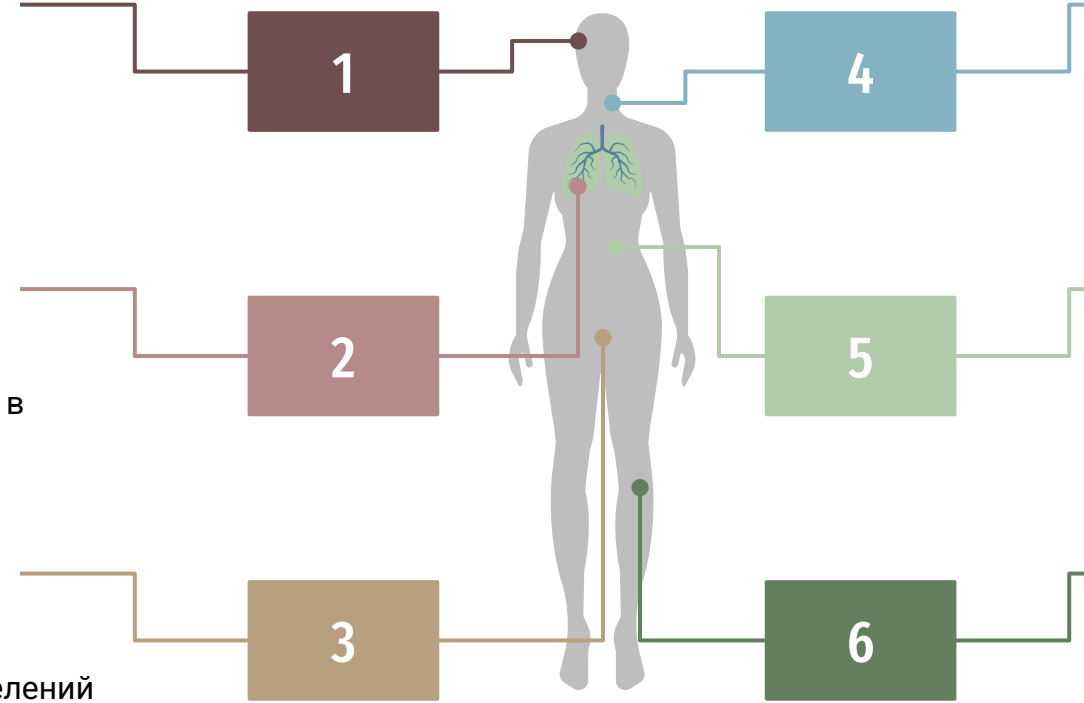
три датасета о респираторных заболеваниях

База данных

единый датасет в формате .db

Анализ данных

вывод распределений и поиск зависимостей



Дашборд

слайд с основными зависимостями

Ключевые факторы

риска заболевания - курение, хронические заболевания, возраст

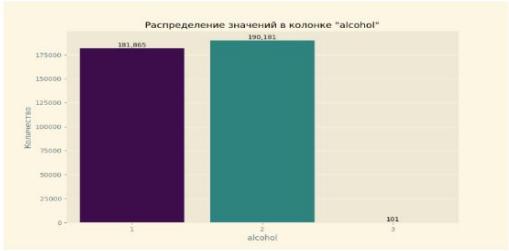
Обучение модели

предсказание риска заболевания с точностью до 83,3%

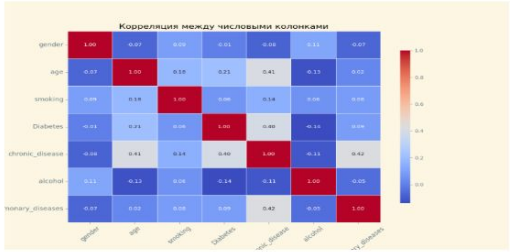
Dashboard: Respiratory Risk Patterns

Age Distribution

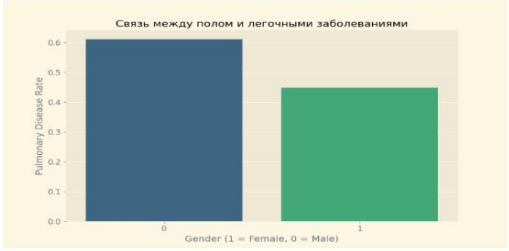
Alcohol Consumption



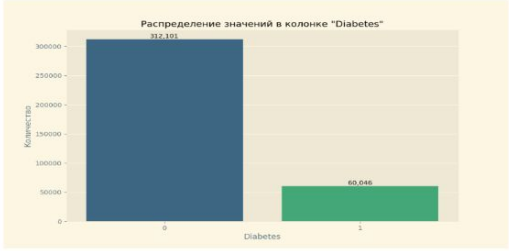
Correlation Heatmap



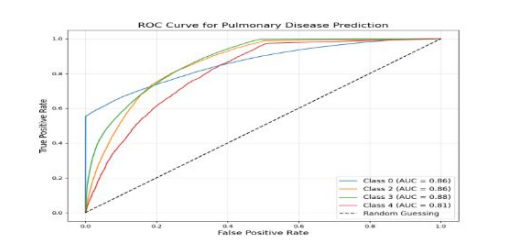
Gender Distribution



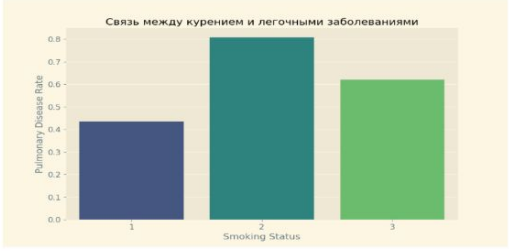
Diabetes Distribution



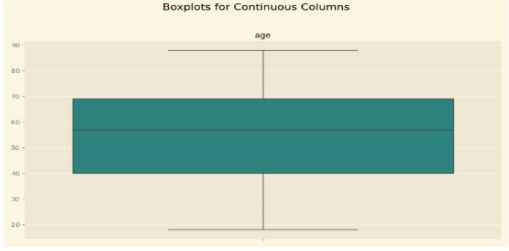
ROC Curve



Smoking Distribution



Boxplots



Feature Importance