Исследование зависмости доли зараженных COVID-19 и его распространения от континента

Серженко Ирина

20 августа 2022 г.

Задача и основная цель проекта
В данном проекте мы выясним, как влияет континент на распространение и протекание коронавируса, рассмотрим зависимость доли зараженных от разничных независимых параметров стран(метод многофакторной линейной регрессии) и классифицируем страны по континентам, опираясь на их независимые признаки(метод k ближайших соседей).

План

- Исследование протекания коронавируса на различных континентах с помощью графиков
- Выявление различных признаков, влияющих на распространение коронавируса по континентам
- Проверить, зависит ли доля зараженных по странам от этих признаков (**Многофакторная линейная регрессия**)
- Классифицировать страны по континентам с помощью отобранных признаков (**метод kNN**)

---

1 Влияет ли континент на распространение ковида на самом деле?

Загрузим данные из файла 'owid-covid-data.csv' и проведем предварительный анализ. Группируя страны по континетам, убедимся в том, что континент влияет на распространение коронавируса. В качестве переменной, выражающей распространение, будет выступать миллионная доля случаев в стране - total cases per million.
Построим график плотности для зависимости миллионной доли случаев от континента(см. Рис. 1)
Заметим, что миллионные доли случаев на некоторых континентах существенно отличаются: например, у почти всех стран в Африке доля достаточно низкая, а в Европе количество стран с долей, превосходящей африканскую, подавляющее большинство.
Так как графики остальных континентов почти идентичны, во 2 блоке мы будем рассматривать только Европу и Африку

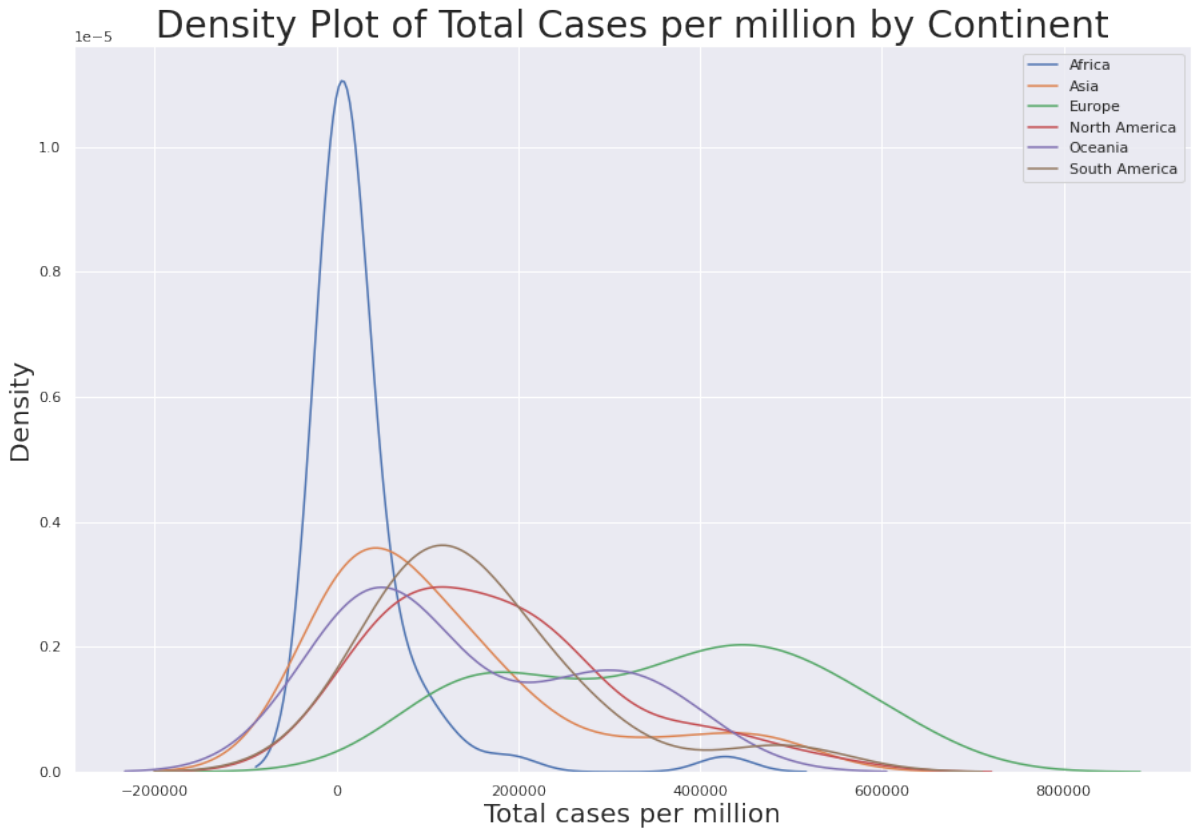


Рис. 1: График распределения стран по миллионной доле зараженных для разных континентов

2 Выявление признаков, влияющих на развитие коронавируса в Европе и Африке

2.1 Влияние вакцинаций на распространение вируса

Для оценки построим графики зависимости доли новых случаев и доли новых вакцинаций от времени для континентов: (см. Рис. 2 и Рис. 3)
Из графиков мы видим, что в Европе напорядок больше вторая волна коронавируса, а прививки европейцев лишь немного превышают Азию, хотя по доле случаев они лидируют гораздо сильнее. Можно предположить, что на распространение ковида оказывают более сильное влияние другие факторы, так как вакцина имеет не самую сильную связь с новыми случаями по континентам.

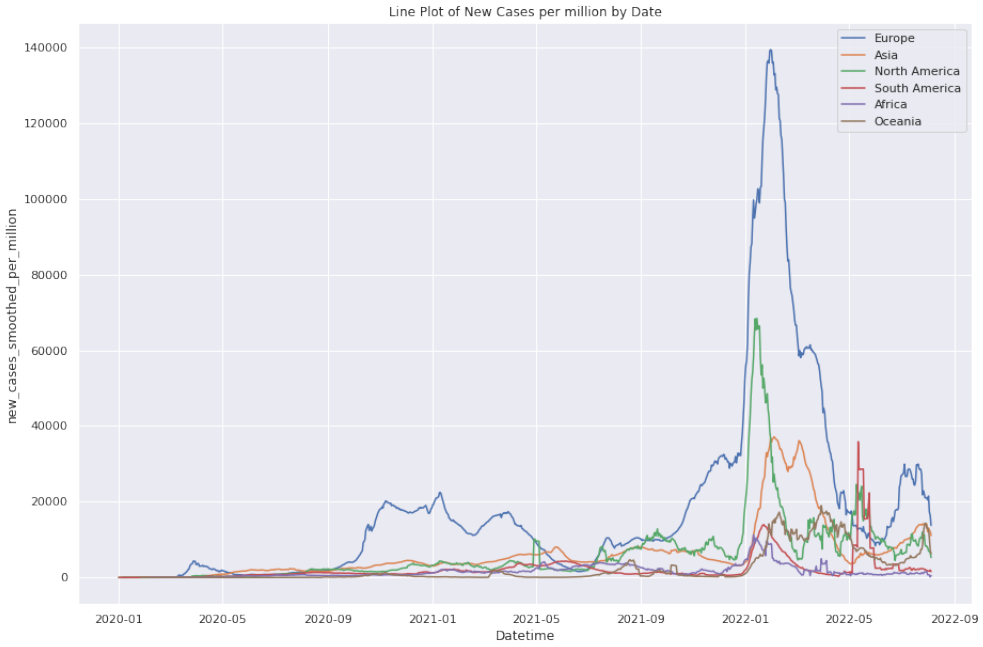


Рис. 2. Зависимость новых случаев от даты

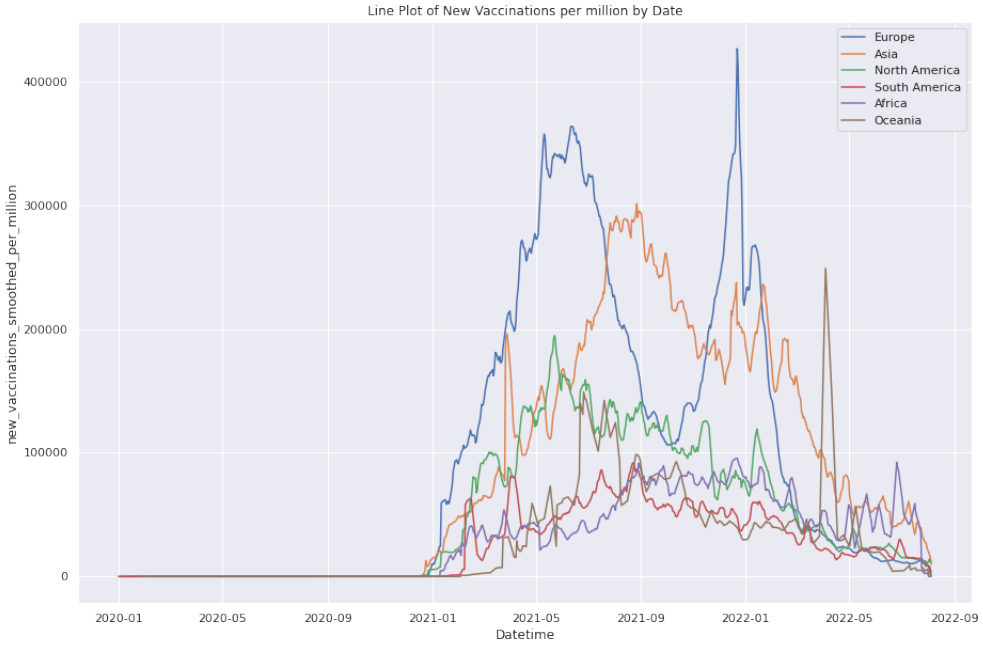
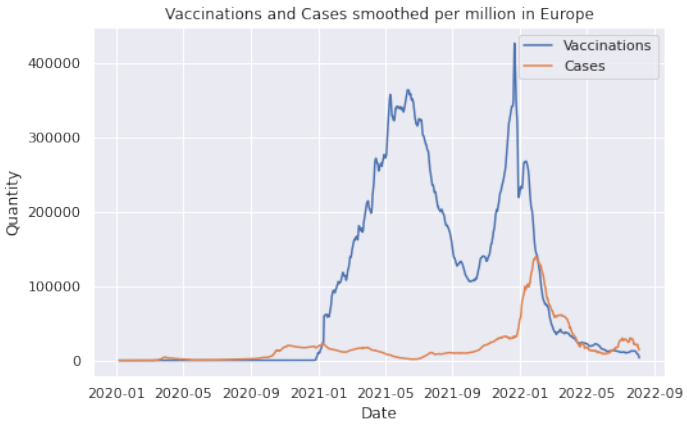


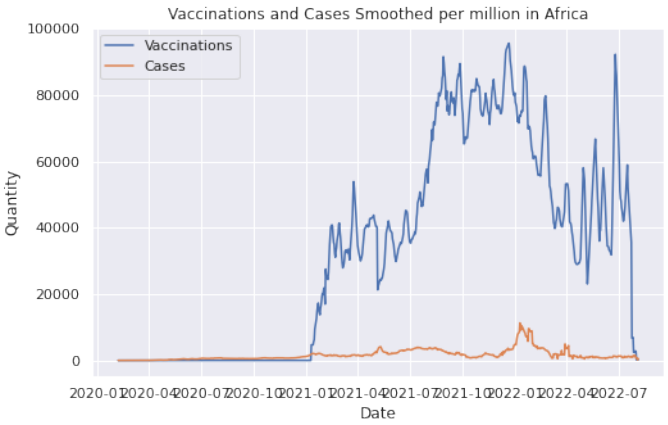
Рис. 3. Зависимость новых вакцинаций от даты

Для лучшего представления зависимости вакцинаций и случаев посмотрим на графики для Европы и Африки: наглядно видно, что в пик 2 волны количество зараженных и количество привитых сравнялось(несмотря на рост заболеваемости, количество привитых в день начало уменьшаться), а в Африке же вакцинации всегда превышали случаи. Вакцинирование определенно

влияет на новые случаи, но можно предположить, что другие параметры(например, строгость карантинных мер либо же степень заинтересованности населения в соблюдении правил защиты от вируса) вкладывают большее значение в долю случаев: проверим это.



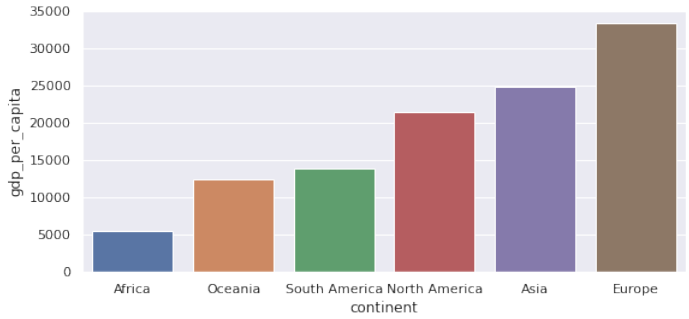
а) для Европы



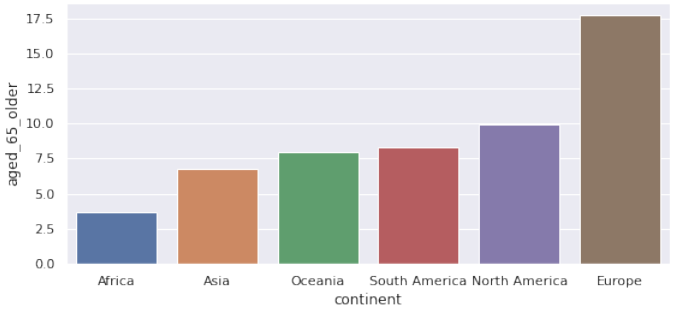
б) для Африки

2.2 Влияние общих признаков

Исследуем влияние признаков, одинаковых для страны любой день наших данных: stringency index(карантинные меры), female/male smokers(процент курящих женщин и мужчин), gdp per capita(ВВП на душу населения), life expectancy(ожидаемая продолжительность жизни), aged 65 older(процент людей старше 65), median age(средний возраст), population density(плотность населения), human development index(индекс человеческого развития) и cardiovasc death rate (смертность от сердечно-сосудистых заболеваний).
Сгруппируем данные по континентам, взяв среднее для каждой величины, и поглядим на результаты: в большинстве признаков, например, ВВП, индекс человеческого развития, процент курящих женщин лидирует Европа, а на последнем месте оказывается Африка.
Гистограммы некоторых примеров на рисунках ниже.



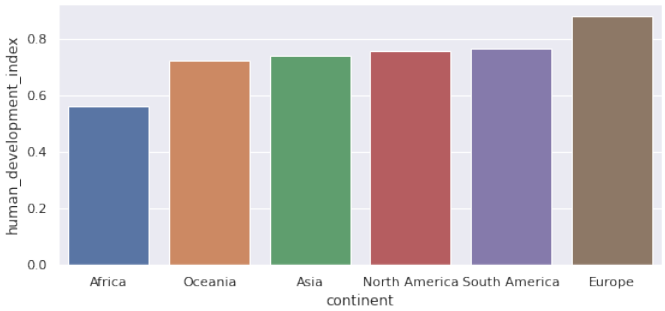
а) ВВП на душу населения



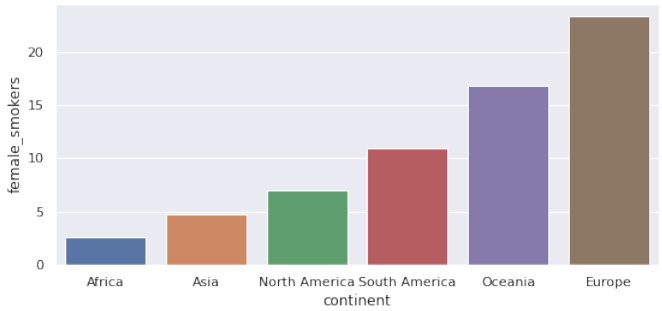
б) процент людей старше 65

Рис. 2: Гистограмма для некоторых общих признаков

Так как данные на различных континентах существенно отличаются, включим их в независимые переменные, оказывающие влияние на миллионную долю зараженных.



в) индекс человеческого развития



г) процент курящих женщин

Рис. 3: Гистограмма для некоторых общих признаков

3 Многофакторная линейная регрессия для миллионной доли случаев

3.1 Африка

В качестве зависимой переменной мы будем использовать миллионную долю всех случаев в каждой стране(total cases per million), а в качестве независимых - общие признаки, такие, как ВВП, качество жизни, процент курящих женщин, количество людей старше 65 лет, и признаки, относящиеся к вирусу напрямую: тысячная доля тестов и сотая доля вакцинаций.
С помощью многофакторной линейной регрессии вычислим коэффициенты на тренировочном наборе данных и посчитаем миллионную долю случаев на тестовом наборе. Сравнение настоящих и предсказанных данных, а также коэффициенты представлены в таблице на рисунке:

	Coef
total_vaccinations_per_hundred	-86.390560
gdp_per_capita	0.545511
female_smokers	1495.713711
aged_65_older	6906.733884
total_tests_per_thousand	101.617440
life_expectancy	-518.888404
male_smokers	7.389090

а) коэффициенты

location	Actual	Predicted
Ghana	5126.118	4177.479651
Namibia	66894.426	58480.728794
Senegal	5187.619	1380.663286
Mauritius	191482.891	146701.635243
Botswana	125838.783	104394.121410
Sierra Leone	918.814	13671.357958
South Africa	67425.542	69240.613818

б) тестовые и прогнозируемые значения

Рис. 4: Данные по линейной регрессии для Африки

Мы видим, что миллионная доля случаев положительно зависит от количества тестов, процента курящих женщин и мужчин, и наиболее сильно от доли людей старше 65. Также отрицательно зависит от качества жизни, и от доли вакцинаций. Оценим качество нашей модели, вызвав функцию r2 score:
R2 score = 0.9122552828905188
Коэффициент получился достаточно близким к единице, соответственно, предложенной нами зависимостью можно описать миллионную долю случаев в разных странах. Несмотря на доста-

точно нетипичный минусовой коэффициент перед долей вакцинаций, модель считает результат с высокой точностью. Также, посмотрим на график распределения тестовой и прогнозируемой величины, чтобы убедиться в точности измерений:
Графики демонстрируют высокую коррелляцию, соответственно, учитывая этот факт и близкое

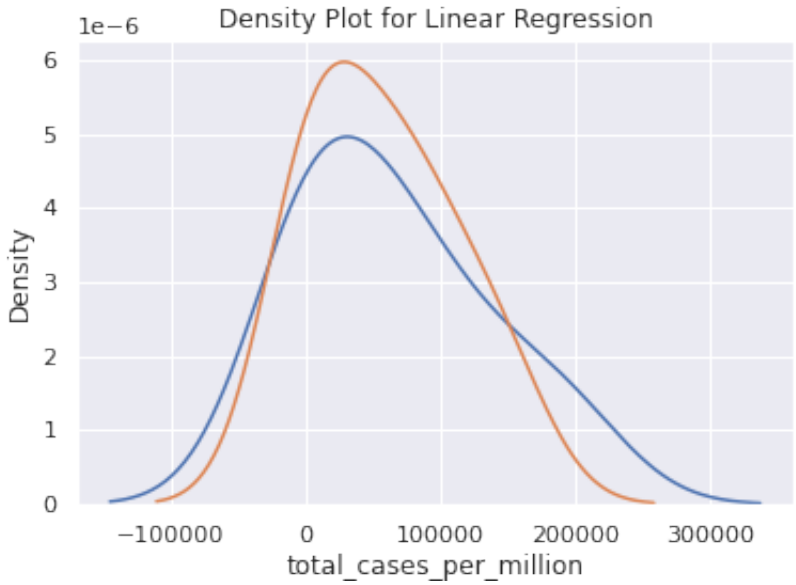


Рис. 5: График распределения тестовой и прогнозируемой величины

к 1.0 значение метрики r2 score, можно сказать, что созданная нами модель применима и верна.

3.2 Европа

В качестве зависимой переменной мы будем использовать миллионную долю всех случаев в каждой стране(total cases per million), а в качестве независимых - общие признаки, такие, как ВВП, качество жизни, процент курящих женщин, количество людей старше 65 лет, и признаки, относящиеся к вирусу напрямую: тысячная доля тестов и сотая доля вакцинаций.
С помощью многофакторной линейной регрессии вычислим коэффициенты на тренировочном наборе данных и посчитаем миллионную долю случаев на тестовом наборе. Сравнение настоящих и предсказанных данных, а также коэффициенты представлены в таблице на рисунке: В

	Coef
total_vaccinations_per_hundred	638.972809
female_smokers	-3485.376072
life_expectancy	-15821.429643
hospital_beds_per_thousand	-16333.090258
stringency_index	-5651.665190
extreme_poverty	-2784.601725
median_age	15897.980626
total_tests_per_thousand	19.842317
human_development_index	778431.065546

а) коэффициенты

location	Actual	Predicted
Belgium	382149.556	316971.732479
Russia	126810.776	227440.201793
Lithuania	427097.616	394569.538904
Norway	269636.376	347171.774020
Croatia	293661.418	246509.449296
Ireland	330411.393	292888.887327

б) тестовые и прогнозируемые значения

Рис. 6: Данные по линейной регрессии для Европы

данной модели **r2 score = 0.5410458919283929**, что является менее точным результатом, чем r2 score для Африки. График распределения говорит об этом же - разница между тестовой и прогнозируемой величиными больше: Данная модель требует уточнения для получения более

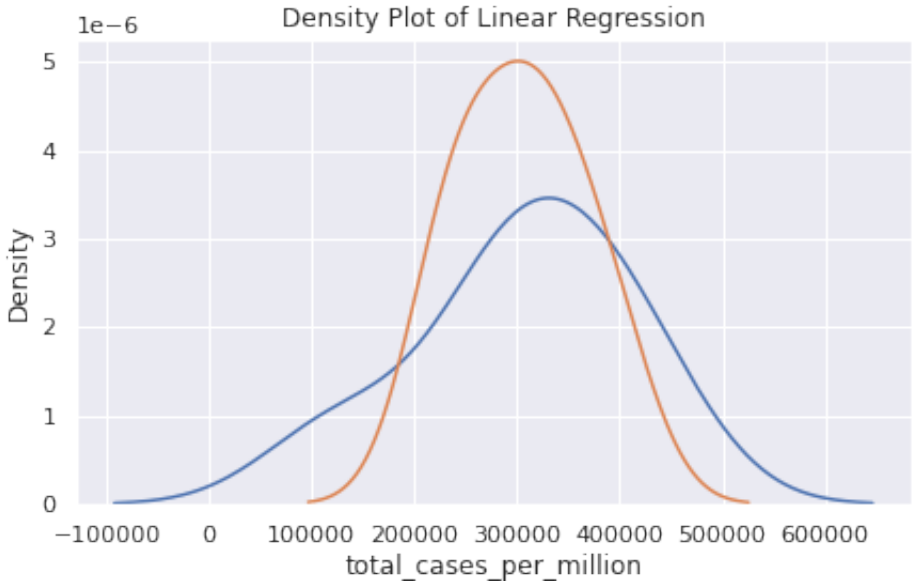


Рис. 7: График распределения тестовой и прогнозируемой величины в Европе

точного результата, например, сбора данных, выступающих в качестве независимых параметров(например, больше данных по госпитализации могло бы изменить ситуацию).

4 Классификация стран по континентам с помощью kNN

В качестве класифицирующих параметров мы будем использовать общие признаки, одинаковые в каждый момент времени: качество жизни, доля курящих женщин и мужчин, ВВП на душу населения, доля людей за чертой бедности, доля людей старше 65, медианный возраст, смертность от сердечно-сосудистых заболеваний, и также признаки, зависяющие от протекания коронавируса в стране: миллионнвя доля случаев, тысячная доля тестов и миллионная доля смертей.

Использовать все признаки в таблице с данными мы не можем, так как в каких-то столбцах заполнено меньше 30 процентов, поэтому количество тестовых и тренировочных записей будет очень маленьким(соответственно ухудшится точность), а остальные не взятые столбцы несут в себе похожую информацию, и не дадут существенных изменений для нашей модели. В полученной таблице с данными для нашего метода 95 различных стран, поэтому в качестве оптимального k для метода kNN мы возьмем

$$k = [\sqrt{N}] + 1 = 10$$

С помощью метода k ближайших соседей обучим модель на тренировочных данных и затем запустим на тестовых. Чтобы оценить качество нашей модели, используем метрику accuracy: **accuracy score = 0.6842105263157895**
Проще говоря, точность модели 68 процентов, что является удовлетворительным результатом. Чтобы визуализировать ошибочно определенные и правильные континенты, нанесем их на график: красным цветом обозначены неправильно определенные точки.

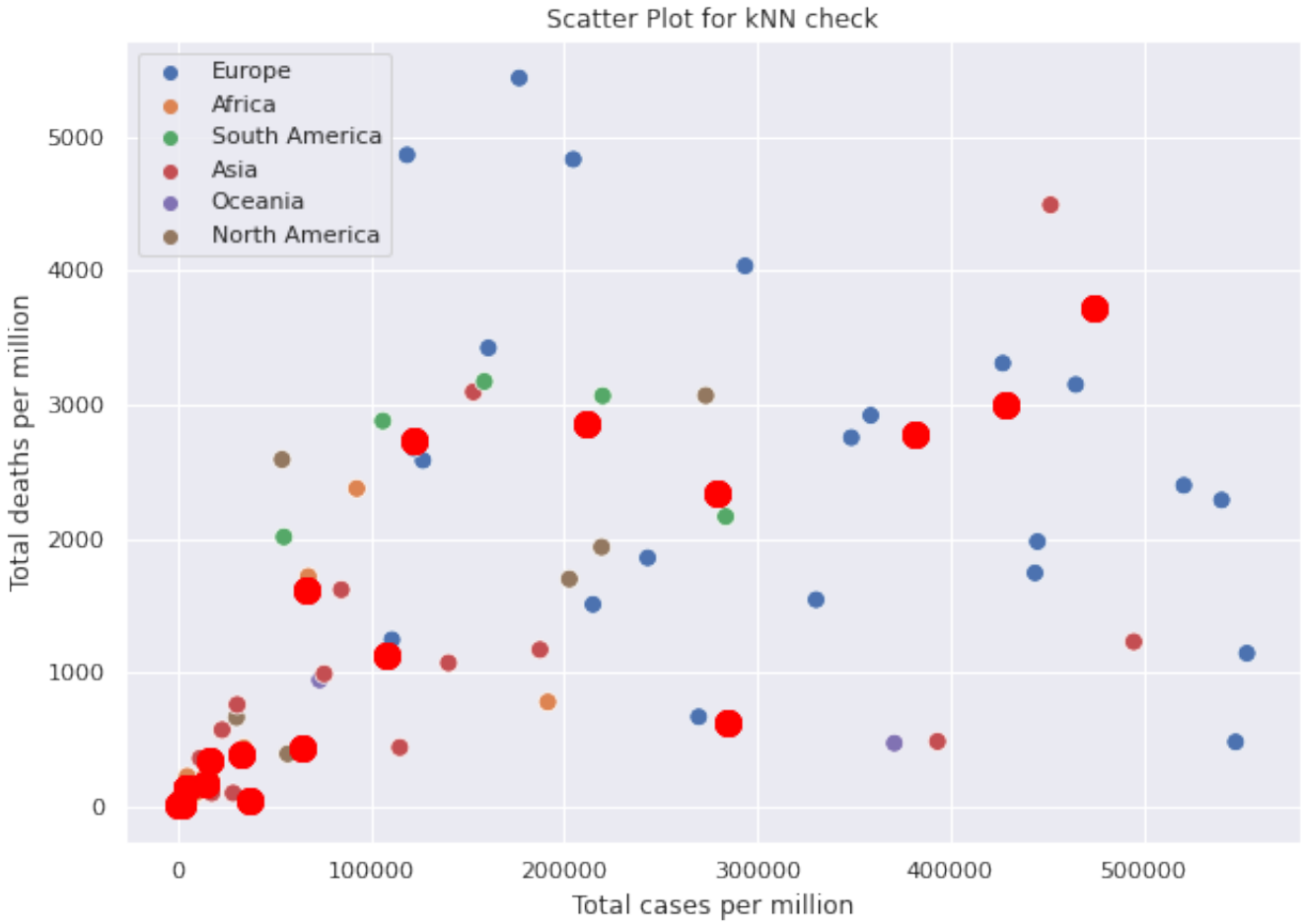


Рис. 8: График распределения тестовой и прогнозируемой величины в Европе

Как мы видим на графике, большинство точек определено верно, а неправильно определенные страны - это в основном страны с маленькой долей заболевших коронавирусом и смертностью. Из этого можно сделать вывод, что распространение ковида по стране действительно зависит от континента, и протекает по разному в каждом из них.
Также, учитывая большинство верно определенных стран, можно сказать, что данная модель с большой точностью определяет континент, и выделенные нами признаки действительно являются характерными для каждого континента в общем случае.

5 Заключение

В ходе данной работы была исследована зависимость протекания коронавируса от континента, были выявлены признаки, оказывающие влияние на миллионную долю коронавируса, и континенты, более всего подверженные данному вирусу. Самой надежной моделью оказалась модель многофакторной линейной регрессии миллионной доли случаев в Африке с коэффициентом детерминации 0, 92, показывающая, что протекание ковида можно выразить через общие характеристики страны, принадлежащей определенному континенту.

Также методом k наиближайших соседей была построена модель классификации стран по континентам, позволяющая определить континент, на котором расположена страна, по данным протекания COVID-19(смертность, доля случаев, доля тестов) и общим данным развития страны(таким, как ВВП на душу населения, доли курящих людей, плотности населения и т.п.) с точностью 0.68

В конце хотелось бы отметить, что распространение и протекание вируса COVID-19 зависит от очень многих факторов, включая признаки, отобранные в этапе нашего исследования, большую роль в характере распространения коронавируса играет именно континент.