

**Исследование влияния множественных
факторов на уровень счастья стран
методами машинного обучения**

Выпускная квалификационная работа Криштаносовой Е.А.

Руководитель: Семендяев Р. Ю.

г. Санкт-Петербург



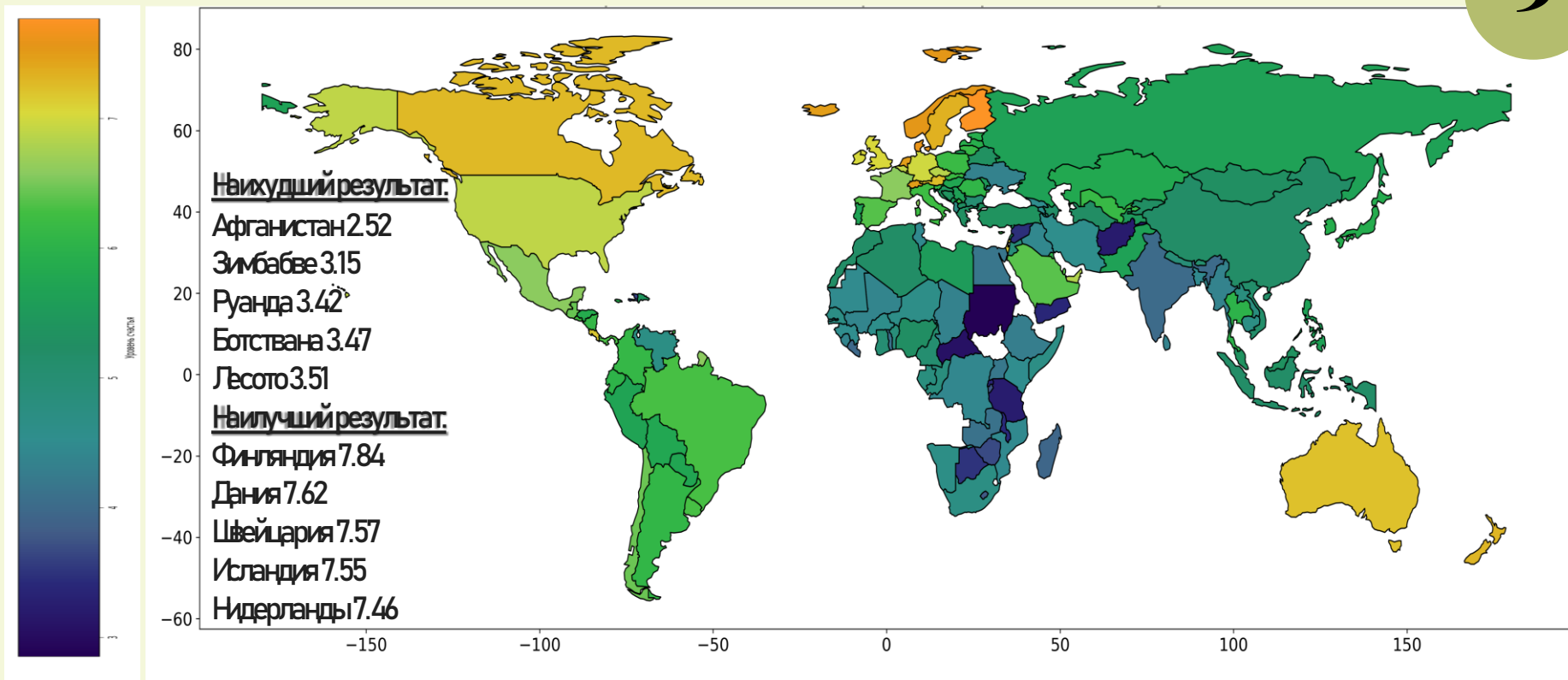
Актуальность

Формирование серьезного **интеллектуального** движения, связанного с попытками исследования счастья было результатом единогласного принятия резолюции **«Счастье: на пути к целостному подходу к развитию»** в 2011 году всеми участниками **Генеральной ассамблеи ООН.**

Сейчас в развитых и развивающихся странах все чаще счастье считается надлежащим мерилom социального прогресса и целью государственной политики.

Уровень счастья в странах мира в 2020 году

3



Результативный признак – «Оценка уровня счастья» взят из «Доклада о мировом счастье 2020 года», которые получены в результате Всемирного опроса Гэллапа. Рейтинги составлены на основе репрезентативных выборок на национальном уровне.

Цель

Поиск и исследование наиболее значимых факторов, влияющих на уровень счастья людей в масштабе страны

4

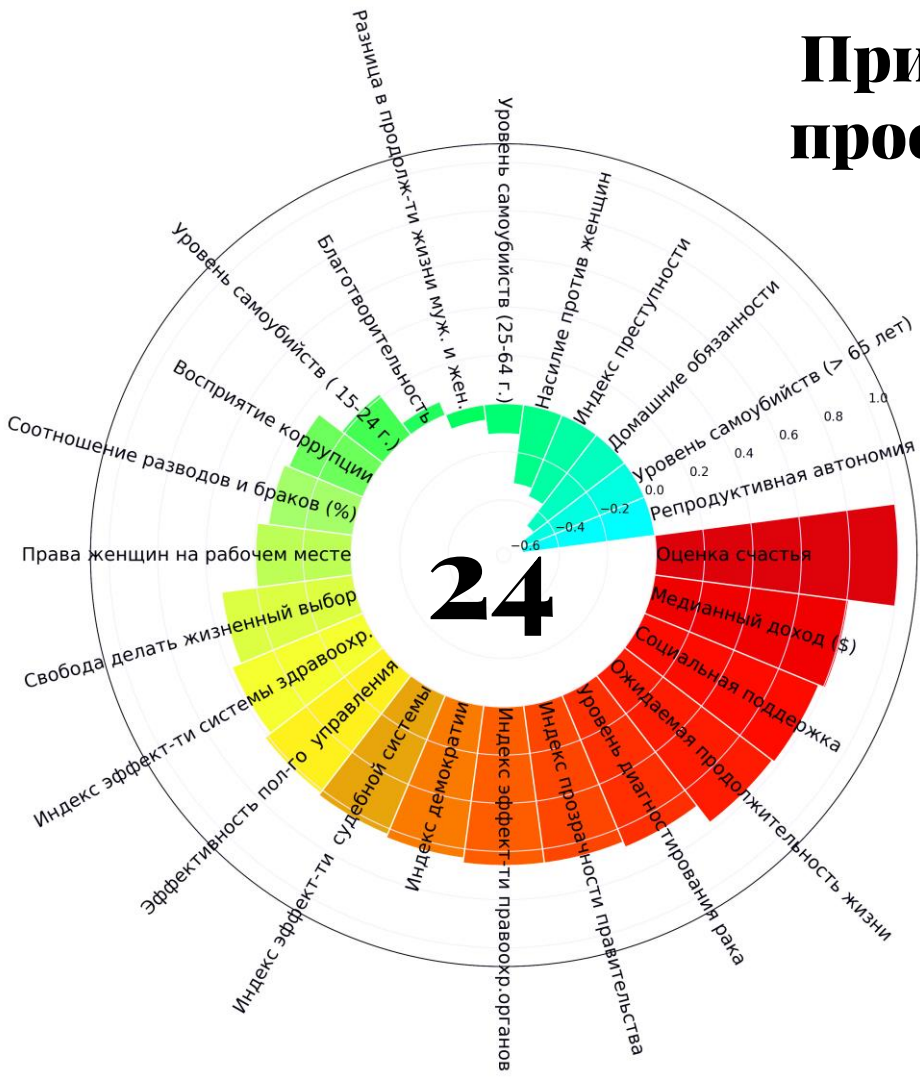
Задачи

- 1** поиск статистических данных из открытых официальных источников информации, проведение разведочного анализа данных
- 2** подбор модели регрессии для выявления наиболее значимых факторов и интерпретации степени влияния этих факторов на уровень счастья стран
- 3** проведение кластерного анализа
- 4** разработка алгоритма практического применения модели кластерного анализа
- 5** апробация практического применения моделей кластеризации и модели регрессии на примере Республики Беларусь

ЗАДАЧА № 1 Поиск статистических данных

- сайт Всемирного банка
- сайт Организации экономического сотрудничества и развития
- сайт Сети решений в области устойчивого развития
- сайт Научно-аналитического подразделения The Economist Group
- сайт Евростата
- сайт Международного всемирного фонда исследования рака
- сайт Департамента по экономическим и социальным вопросам ООН
- сайт Numbeo Краудсорсинговой глобальной базы Numbeo
- сайт Центра по исследованию коррупции и организованной преступности
- сайт Всемирной организации здравоохранения

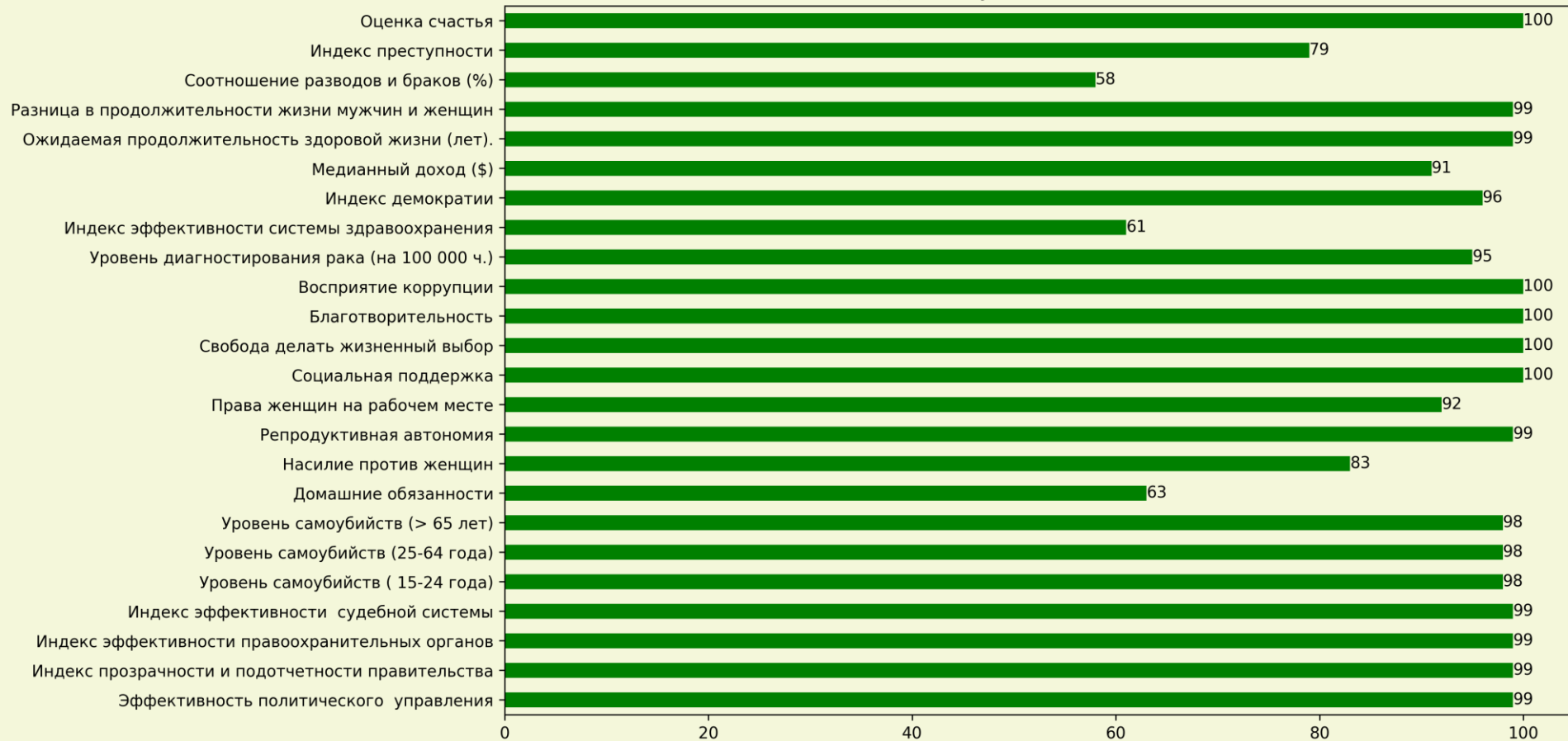
Признаковое пространство



Наличие пропусков в данных

7

Доля ненулевых значений (%)



Стратегия заполнения пропусков:

8

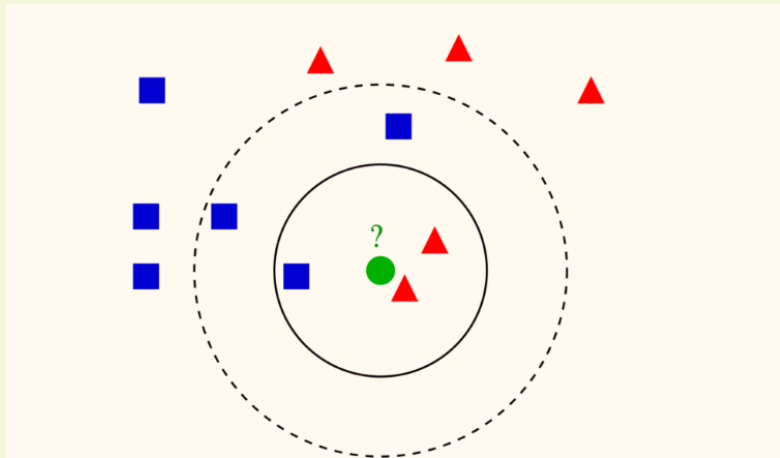
Метод MICE – это многомерный способ с помощью цепных уравнений

Возраст	Опыт	ЗП
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
?	11	130

линейная регрессия →

Возраст	Опыт	ЗП
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
34,99	11	130

Метод k-ближайших соседей



1. Заполняются пропуски многомерным методом заполнения пропусков (MICE)

2. Заполняются пропуски методом k-ближайших соседей

3. Строится модель линейной регрессии по двум наборам и выбирается лучший метод по метрике MSE

Method: KNNImputer, MSE: 0.1718

Method: MICE_ MSE: 0.1766

План действий по отбору признаков:

1. Отбор методом информационного прироста (Information Gain)

Выбираем признаки с наибольшим информационным приростом в контексте целевой переменной

2. Дисперсионный анализ (метод ANOVA)

Выбираем признаки, для которых гипотеза о равенстве дисперсии отвергается (т.е. p -value меньше 0.05)

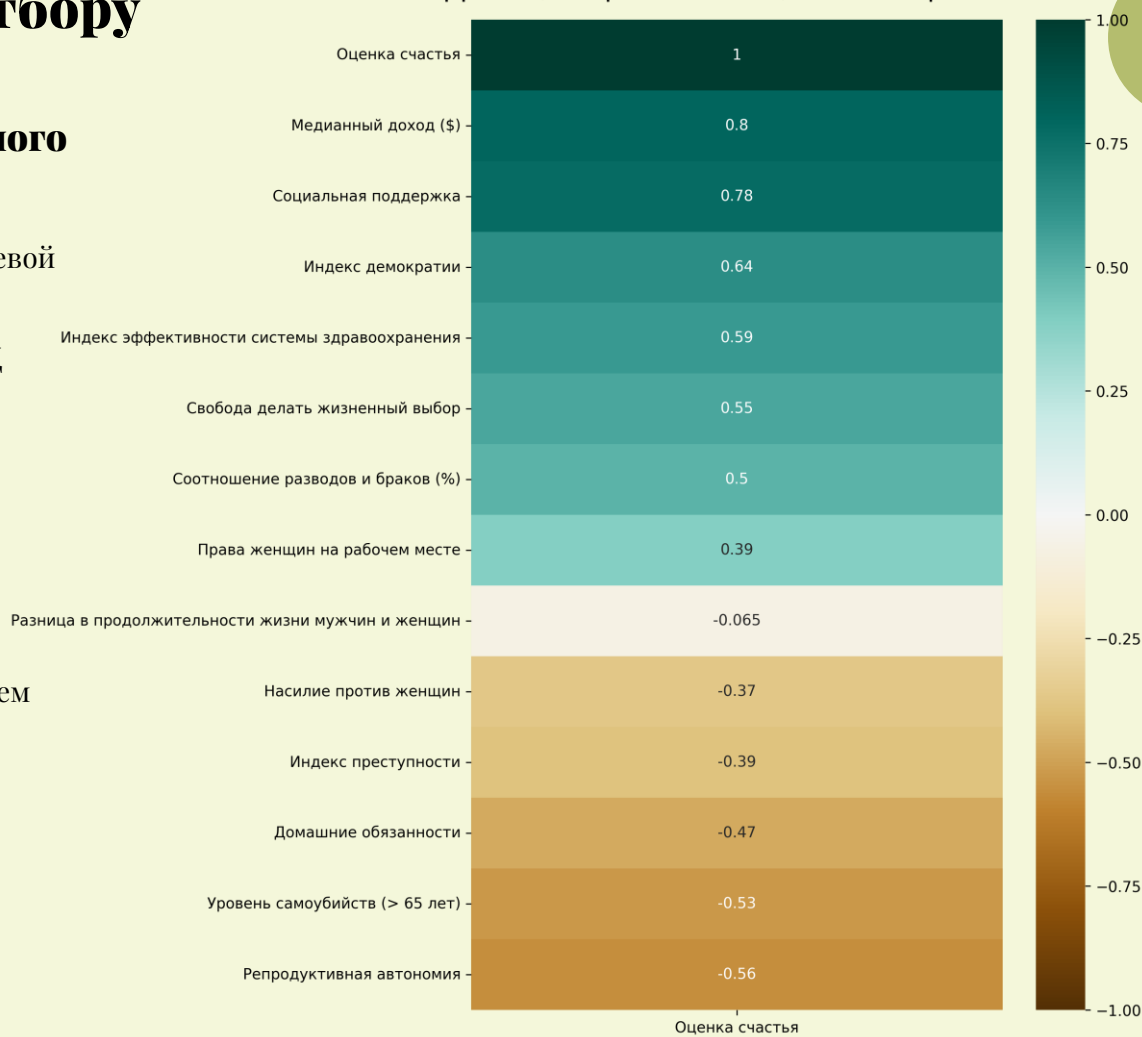
3. Отбор признаков методом случайных деревьев (ExtraTreesRegressor)

Выбираем признаки с максимальным значением `feature_importances`.

4. Корреляция с целевой переменной

Выбираем признаки, которые имеют наиболее тесную корреляцию с результативным признаком

Корреляция признаков с целевой переменной



9

Отбор признаков

Признак	Причина сокращения
<ul style="list-style-type: none">• Эффективность политического управления• Индекс прозрачности и подотчетности правительства• Индекс эффективности правоохранительных органов	По причине сильной корреляции между собой, из этой группы признаков остался «Индекс эффективности судебной системы»
<ul style="list-style-type: none">• Благотворительность• Уровень самоубийств (15 - 24 года)• Уровень самоубийств (24- 65 лет)	Были признаны малозначительными по результатам корреляционного анализа и методов отбора признаков
<ul style="list-style-type: none">• Уровень диагностирования рака (на 100 000 ч.)• Ожидаемая продолжительность здоровой жизни (лет)• Индекс эффективности судебной системы	По причине сильной корреляции с остальными признаками во избежание проблемы мультиколлинеарности

Признаковое пространство для модели регрессии

11

Уровень самоубийств (> 65 лет)	1.0	0.1	0.3	0.5	-0.2	-0.5	-0.2	0.1	-0.3	-0.4	-0.2	0.3	0.1	-0.5
Домашние обязанности	0.1	1.0	0.2	0.2	-0.4	-0.6	-0.3	-0.3	-0.3	-0.4	-0.1	-0.3	0.1	-0.5
Насилие против женщин	0.3	0.2	1.0	0.4	-0.3	-0.4	-0.2	-0.1	-0.3	-0.3	-0.1	-0.0	0.4	-0.3
Репродуктивная автономия	0.5	0.2	0.4	1.0	-0.3	-0.6	-0.3	-0.2	-0.5	-0.5	-0.3	-0.3	0.2	-0.6
Права женщин на рабочем месте	-0.2	-0.4	-0.3	-0.3	1.0	0.5	0.3	0.1	0.4	0.2	0.3	0.2	0.1	0.4
Социальная поддержка	-0.5	-0.6	-0.4	-0.6	0.5	1.0	0.4	0.5	0.6	0.6	0.2	0.4	-0.3	0.8
Свобода делать жизненный выбор	-0.2	-0.3	-0.2	-0.3	0.3	0.4	1.0	0.4	0.4	0.4	-0.2	0.1	-0.2	0.5
Индекс эффективности системы здравоохранения	0.1	-0.3	-0.1	-0.2	0.1	0.5	0.4	1.0	0.5	0.6	-0.1	0.4	-0.4	0.6
Индекс демократии	-0.3	-0.3	-0.3	-0.5	0.4	0.6	0.4	0.5	1.0	0.7	-0.0	0.4	-0.2	0.6
Медианный доход (\$)	-0.4	-0.4	-0.3	-0.5	0.2	0.6	0.4	0.6	0.7	1.0	-0.1	0.5	-0.5	0.8
Разница в продолжительности жизни мужчин и женщин	-0.2	-0.1	-0.1	-0.3	0.3	0.2	-0.2	-0.1	-0.0	-0.1	1.0	0.0	0.1	-0.1
Соотношение разводов и браков (%)	0.3	-0.3	-0.0	-0.3	0.2	0.4	0.1	0.4	0.4	0.5	0.0	1.0	-0.2	0.4
Индекс преступности	0.1	0.1	0.4	0.2	0.1	-0.3	-0.2	-0.4	-0.2	-0.5	0.1	-0.2	1.0	-0.4
Оценка счастья	-0.5	-0.5	-0.3	-0.6	0.4	0.8	0.5	0.6	0.6	0.8	-0.1	0.4	-0.4	1.0

Variance inflation factor

- VIF Уровень самоубийств (> 65 лет): 1.87
- VIF Домашние обязанности: 1.79
- VIF Насилие против женщин: 1.46
- VIF Репродуктивная автономия: 2.13
- VIF Права женщин на рабочем месте: 2.09
- VIF Социальная поддержка: 3.01

- VIF Свобода делать жизненный выбор: 1.58
- VIF Индекс эффективности системы здравоохранения: 2.14
- VIF Индекс демократии: 2.58
- VIF Медианный доход (\$): 3.69
- VIF Разница в продолжительности жизни мужчин и женщин: 1.68
- VIF Соотношение разводов и браков (%): 1.93
- VIF Индекс преступности: 1.66

ПОДБОР МОДЕЛИ РЕГРЕССИИ

12

Были применены:

Линейные модели:

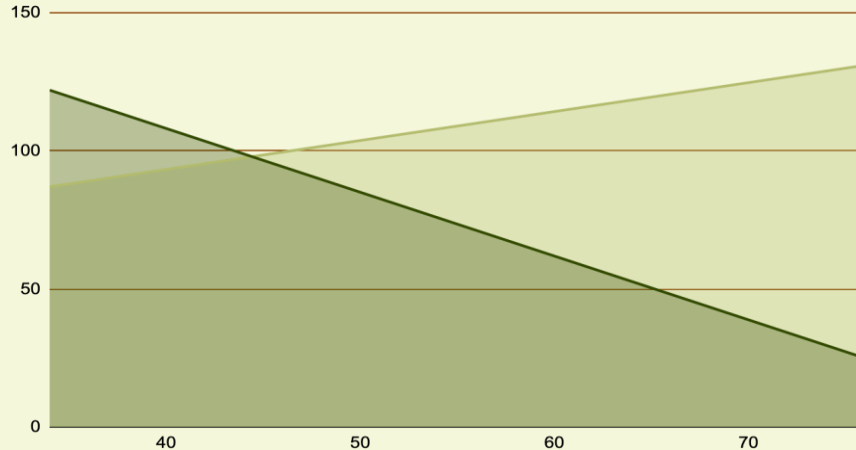
- модель LinearRegression
- модель Ridge
- модель Lasso
- модель ElasticNet
- метод опорных векторов SVR

Нелинейные модели:

- метод kNN KNeighborsRegressor
- модель решающего дерева DecisionTreeRegressor

Ансамблевые методы:

- метод дополнительных деревьев ExtraTreesRegressor
- метод GradientBoostingRegressor
- метод XGBRegressor



1

Разбиение данных на тестовую и тренировочную выборки

2

Настройка гиперпараметров для каждой из моделей с помощью GridSearchCV;

3

Обучение моделей с подобранными гиперпараметрами, построение сравнительных графиков с помощью matplotlib на основе полученных данных;

4

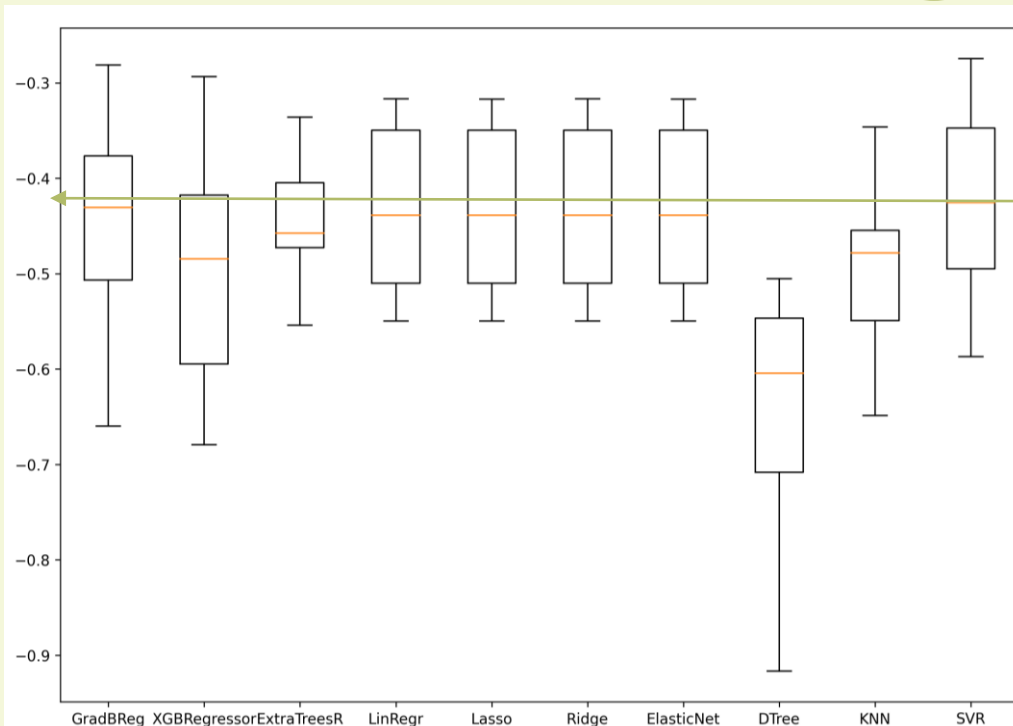
Выбор наиболее оптимальной модели с наименьшей ошибкой по выбранной метрике качества

Сравнение и выбор модели регрессии

13

Ошибки по моделям:

- SVR: -0.388
- ElasticNet: -0.398
- Ridge: -0.399
- LinRegr: -0.399
- Lasso: -0.402
- GradBReg: -0.413
- XGBRegressor: -0.421
- ExtraTreesR: -0.443
- KNN: -0.460
- DTree: -0.525

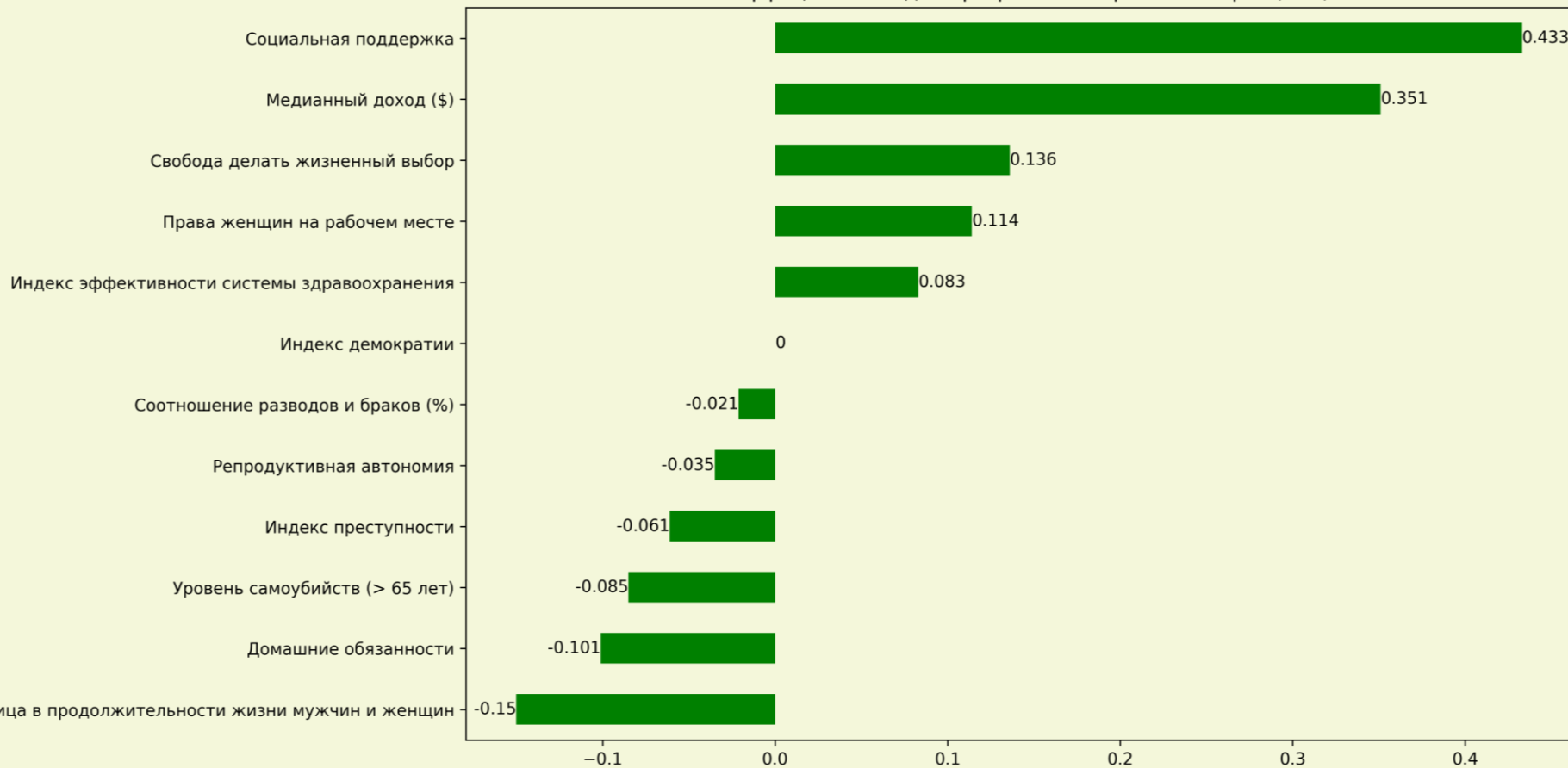


Лучше всего зарекомендовала себя **модель опорных векторов SVR**. Средняя абсолютная ошибка равна 0.388, что является неплохим результатом.

Значимость факторов по коэффициентам модели регрессии

14

Коэффициенты модели регрессии опорных векторов (SVR)



Кластеризация стран

15

Оценка результатов кластеризации по выбранным признакам и сравнение среднего уровня счастья в кластерах позволит добавить объективности в исследование счастья в разных странах мира.

В данной работе продемонстрированы три способа кластерного анализа:

1. Иерархический

2. Кластерный анализ методом k-средних

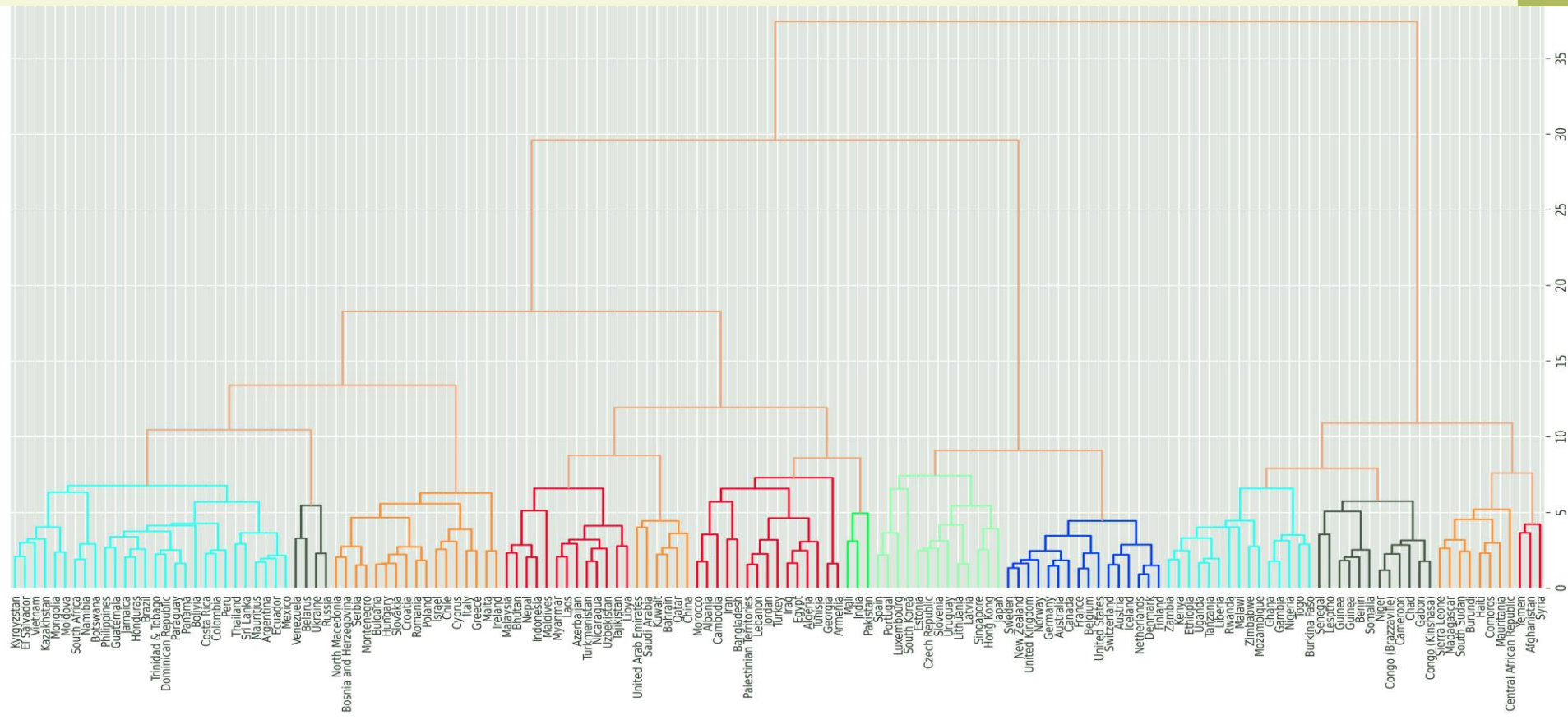
3. Спектральный кластерный анализ



А что если есть субъективность в понимании счастья в странах с разными культурными традициями?

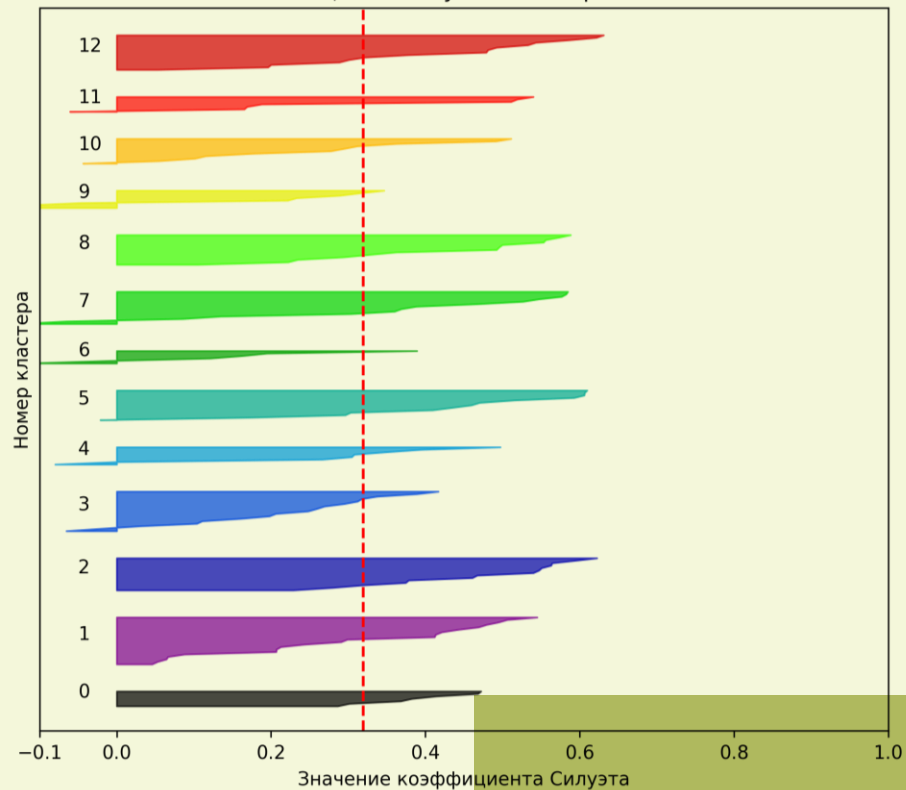
Иерархический кластерный анализ

Полученная с помощью функции linkage дендрограмма говорит о том, что на основе имеющихся признаков данные целесообразно разбить на 13 кластеров

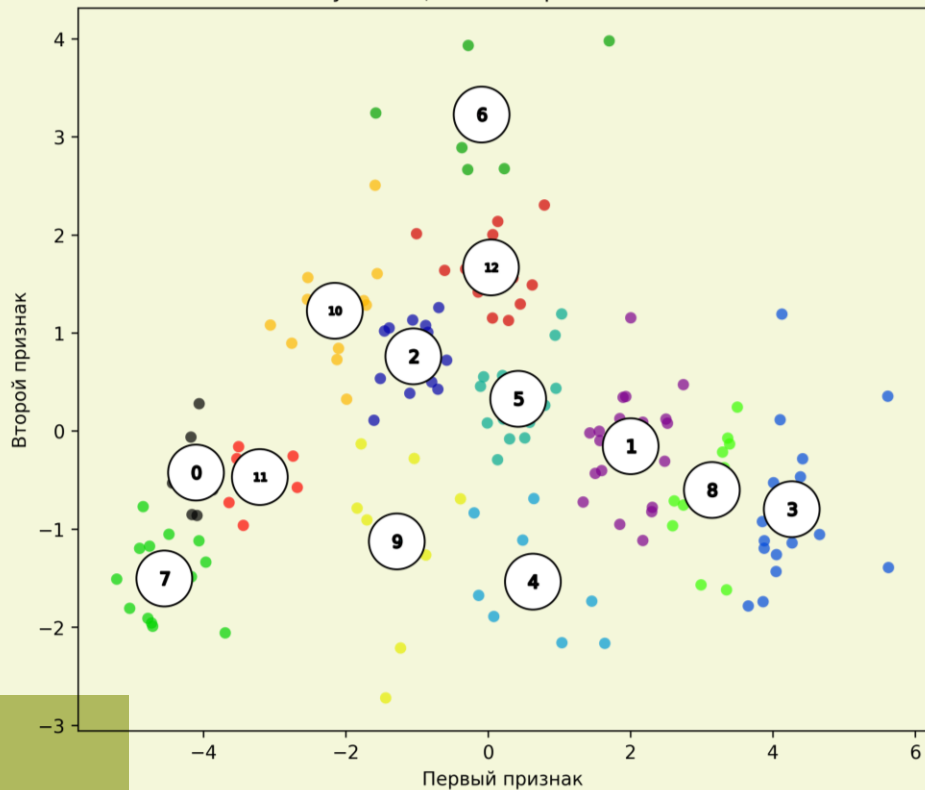


Визуализация кластеризации методом k-mean

Оценка Силуэта кластеров



Визуализация кластерного анализа



Результат кластеризации стран методом k-mean

18

№ кластера по рангу	Средняя оценка счастья	Названия стран	Название кластера
1	7,12	Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Iceland, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, United States	Процветающие нации
2	6,33	Chile, Czech Republic, Estonia, Ireland, Israel, Italy, Japan, Slovenia, Cyprus, Hong Kong, Malta, Singapore, Uruguay	Восходящие гиганты
3	6,05	Colombia, Costa Rica, Mexico, Argentina, Bolivia, Ecuador, Guatemala, Jamaica, Malaysia, Mauritius, Panama, Paraguay, Peru, Philippines, Thailand	Пост колониальные многообещающие земли
4	5,63	Greece, Hungary, Poland, Slovakia, Bulgaria, Croatia, North Macedonia, Montenegro, Romania, Serbia	Посткоммунистические амбициозные страны
5	5,55	Bahrain, Bhutan, China, Jordan, Kuwait, Maldives, Qatar, Rwanda, Saudi Arabia, United Arab Emirates	Изобретательные царства
6	5,49	Latvia, Lithuania, Belarus, Moldova, Russia, Ukraine	Переходные экономики
7	4,94	South Korea, Turkey, Armenia, Bosnia and Herzegovina, Georgia, Lebanon, Palestinian Territories, Sri Lanka, Tunisia	Плюралистические провинции
8	5,47	Brazil, Dominican Republic, El Salvador, Honduras, Kazakhstan, Kyrgyzstan, Mongolia, Namibia, South Africa, Trinidad & Tobago, Venezuela, Vietnam	Суровые владения
9	5,30	Azerbaijan, Indonesia, Laos, Libya, Myanmar, Nepal, Nicaragua, Tajikistan, Turkmenistan, Uzbekistan	Растущие рынки
10	4,68	Albania, Algeria, Bangladesh, Cambodia, Egypt, India, Iran, Iraq, Mali, Morocco, Pakistan	Густонаселенные провинции
11	4,14	Botswana, Burkina Faso, Ethiopia, Gambia, Ghana, Kenya, Lesotho, Malawi, Mozambique, Senegal, Tanzania, Uganda, Zambia, Zimbabwe	Сельские районы
12	4,56	Benin, Burundi, Cameroon, Congo (Brazzaville), Congo (Kinshasa), Guinea, Gabon, Guinea, Liberia, Niger, Nigeria, Sierra Leone, Somalia, Togo	Обедневшие нации
13	3,63	Afghanistan, Central African Republic, Chad, Comoros, Haiti, Madagascar, Mauritania, South Sudan, Syria, Yemen	Истерзанные войной территории

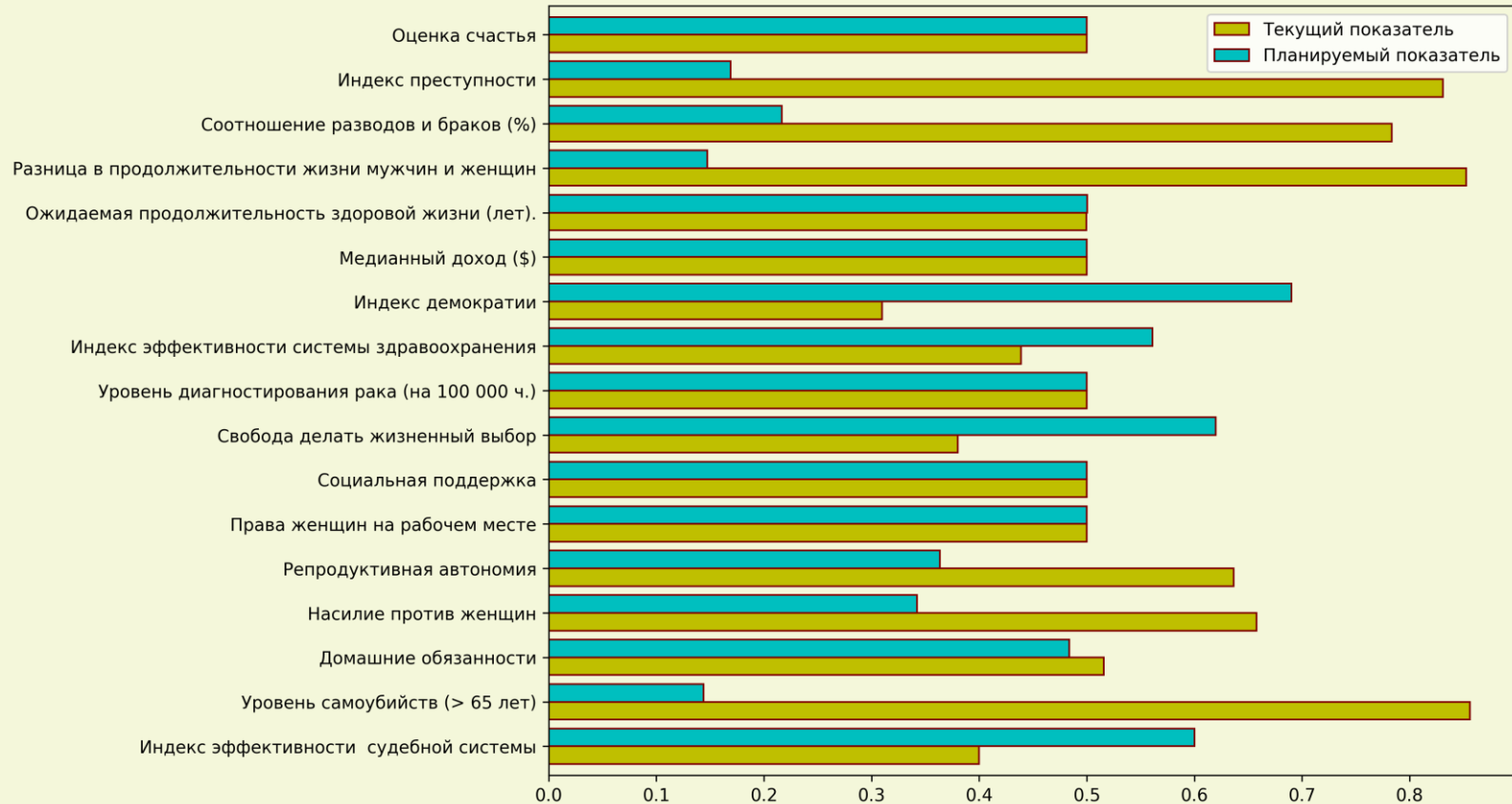
Какие проблемы придется преодолеть Республике Беларусь на пути к более высокому социальному благополучию и уровню счастья?

Наименования признаков	№4	№5	Текущий для РБ	Рек-й показатель	Показатель в %
Индекс эффективности судебной системы	4,5	3	3	4,5	50%
Уровень самоубийств (> 65 лет)	21,66	21,27	126,59	21,27	-83%
Домашние обязанности	1,82	1,81	1,93	1,81	-6%
Насилие против женщин	13	16,96	25	13	-48%
Репродуктивная автономия	9	4	7	4	-43%
Права женщин на рабочем месте	24	1	46	46	0%
Социальная поддержка	1,16	0,71	1,47	1,47	0%
Свобода делать жизненный выбор	0,07	0,38	0,24	0,38	58%
Уровень диагностирования рака (на 100 000 ч.)	366,3	48,6	442,5	442,5	0%
Индекс эффективности системы здравоохранения	52,4	58,9	46,1	58,9	28%
Индекс демократии	5,77	2,08	2,59	5,77	123%
Медианный доход (\$)	4279	925	7359	7359	0%
Ожидаемая продолжительность здоровой жизни (лет).	66	60	66	66	0%
Разница в продолжительности жизни мужчин и женщин	4,37	1,76	10,17	1,76	-83%
Соотношение разводов и браков (%)	12,31	24,66	44,57	12,31	-72%
Индекс преступности	24,59	12,13	59,58	12,13	-80%

Апробация применения модели кластеризации на примере Республики Беларусь




20

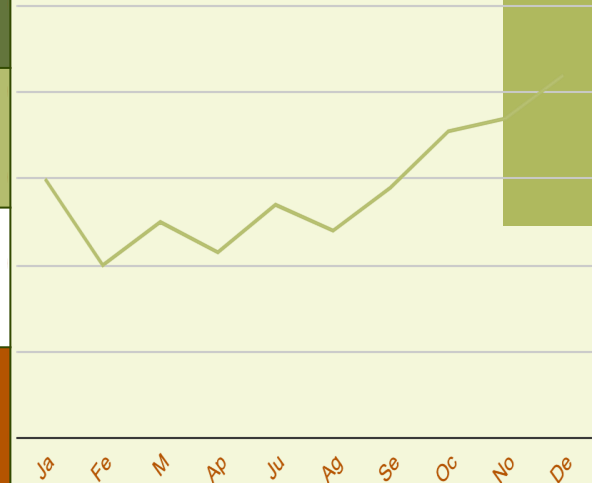
Сравнение текущих и расчетных (для перехода в вышестоящий кластер) показателей для РБ



Апробация модели регрессии

21

	Показатель	Значение
	Реальный уровень счастья в 2020 году в Республике Беларусь	5,3
	Рассчитанный уровень счастья по модели	5,5
	Рассчитанный уровень счастья по модели с прогнозными данными	5,9



Результаты применения модели кластеризации

Изначально номер кластера Республики Беларусь – номер шесть. Рекомендуемые показатели мы внесли в качестве новых аргументов в метод predict модели кластеризации Kmean и получили новое расчетное значение кластера для РБ – кластер номер четыре.

Выводы

Были выявлены наиболее значимые факторы, влияющие на уровень счастья в странах мира по модели регрессии:

- 1. Социальная поддержка**
- 2. Медианный доход (\$)**
- 3. Свобода делать жизненный выбор**
- 4. Разница в продолжительности жизни мужчин и женщин**
- 5. Права женщин**
- 6. Домашние обязанности**

Проведена кластеризация стран для проверки объективности имеющейся целевой переменной (уровня счастья) и определения тактики повышения уровня счастья для каждой отдельно взятой страны

Проведена апробация моделей кластеризации и модели регрессии на примере Республики Беларусь. Таким образом, были определены показатели, которых необходимо достичь для повышения уровня счастья в этой стране и рассчитан предполагаемый уровень счастья при условии достижения этих показателей.



Спасибо за внимание

Исследование влияния множественных факторов на уровень счастья стран методами машинного обучения

Выпускная квалификационная работа Криштаносовой Е.А.

Руководитель: Семендяев Р. Ю.

г. Санкт-Петербург