Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчет

по итоговой работе

«**Анализ предпринимательской активности в социокультурном контексте**»

по дисциплине «Анализ культурных данных»

Студент: Малаев С.Г.

Факультет: ПИн

Группа: К34422

Преподаватель: Коцюба И. Ю.

Санкт-Петербург,

2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Предпринимательство тесно связано с социокультурными условиями общества. Направление предпринимательской деятельности в контексте культуры означает различия в типах и сферах бизнеса, обусловленные культурными факторами, от восприятия предпринимательства обществом до преобладающих отраслей, уровня инноваций и социальной направленности бизнеса.

К примеру, культурные установки могут влиять на готовность людей начинать свое дело, по данным глобального мониторинга GEM, почти половина опрошенных в мире, 49% в 2024 году, заявили, что страх неудачи удерживает их от открытия бизнеса, против 44% пятью годами ранее.

С другой стороны, за последние десятилетия растет участие женщин в предпринимательстве. В 2023 году стартап активность среди женщин достигла ~10,4%, одна из десяти женщин начинала новый бизнес, тогда как среди мужчин около 12%, один из восьми мужчин.

Тем не менее, в странах с высоким доходом фиксируются самые низкие показатели женского предпринимательства и наибольшей гендерный разрыв.

Эти факты демонстрируют, насколько важно изучать предпринимательскую активность с учетом социокультурных факторов.

1 Цель исследования

Главная задача – проанализировать на реальных данных, как социокультурный контекст влияет на предпринимательскую активность, и предложить соответствующие математические модели анализа.

Будут подобраны релевантные датасеты, отражающие разнообразие предпринимательства в разных культурах, и применены проблемно-ориентированные подходы анализа.

Исследование последовательно охватывает:

– выбор данных,

– применение различных практик анализа,

– построение математических моделей,

– интерпретацию полученных результатов.

2 Подбор данных

Для анализа были выбраны открытые данные с платформы Kaggle. Основным источником стал датасет “Women Entrepreneurship and Labor Force”, содержащий показатели для 50+ стран мира. Данные получены из отчета о Глобальном индексе предпринимательства (Global Entrepreneurship Index) и Индексе женского предпринимательства (Women Entrepreneurship Index) за 2015 год, опубликованного в открытом доступе.

Выбранный набор данных важен для понимания влияния культуры через гендерные различия и макроэкономический контекст, которые часто обусловлены историческими и культурными особенностями общества.

Дополнительно, для иллюстрации микроуровня предпринимательства, был рассмотрен датасет “Entrepreneurial Competency in University Students”. Этот набор данных содержит сведения об 219 студентах в Индии, включая 16 характеристик и целевую переменную, отражающую вероятность того, что студент собирается стать предпринимателем. Данный набор служит примером индивидуального уровня анализа, позволяя изучить, какие личные и культурные факторы влияют на предпринимательские намерения. Он полезен для демонстрации классификационных моделей, по сути, это задача предсказания, где определяется станет ли студент предпринимателем на основе его компетенций.

3 Предварительная обработка данных

Перед проведением анализа данных были применены стандартные практики их обработки:

– очистка данных: датасеты были проверены на наличие пропущенных значений, аномалий и ошибок. Для датасета студентов также проверялась целостность ответов на опрос и консистентность кодирования категорий,

– нормализация показателей: показатели разных шкал были масштабированы по необходимости. В частности, для корректного применения методов кластеризации и факторного анализа данные нормализуются, предотвращая доминирование переменных с большим разбросом над другими.

Ключевые шаги выполненные в работе:

1) удаление нерелевантных признаков,

2) обработка пропусков,

3) кодирование категориальных признаков,

4) масштабирование численных признаков.

На первом этапе работы с данными Women Entrepreneurship Index (WEI) был создан отдельный рабочий набор данных, в котором из исходного датасета были удалены все вспомогательные и неинформативные столбцы, такие как порядковый номер, информация о валюте и принадлежности к Европейскому союзу. Для переменной “Level of development” была выполнена процедура кодирования категориальных данных в числовой формат с помощью LabelEncoder, что позволило использовать данный признак в дальнейших числовых методах анализа.

Столбец “Country” был временно сохранен для последующего восстановления меток, а затем удален, чтобы предотвратить влияние на результат кластеризации и факторного анализа. Особое внимание было уделено проверке данных на наличие пропусков и аномалий, однако их не было выявлено. Ключевые количественными показателями являются индексы предпринимательства среди женщин и в целом, уровень инфляции и участие

женщин в рабочей силе. Они были подвергнуты стандартизации методом “z-преобразования”.

Для датасета по предпринимательским компетенциям студентов также была выполнена очистка. Из набора удалены текстовые и факультативные признаки, не используемые в построении моделей, в том числе информация о городе, индивидуальных проектах и причинах отказа от предпринимательства. Пропуски данных были проверены и устранены на этапе отбора признаков. Категориальные переменные, такие как сектор образования, пол, факт наличия вдохновляющих примеров в окружении и наличие ментальных расстройств, были преобразованы в числовой вид с помощью процедуры кодирования. Далее датасет был разделен на матрицу признаков и целевую переменную, отражающую склонность к предпринимательству. Для восьми ключевых количественных характеристик была проведена стандартизация значений.

4 Первичная исследовательская аналитика

Перед построением сложных моделей был проведен всесторонний разведочный анализ данных, целью которого являлось первичное выявление структурных особенностей, базовых зависимостей и аномалий в распределениях ключевых переменных.

В рамках EDA для каждого датасета были вычислены основные описательные статистики, а также построены наглядные графики.

Для странового датасета были рассчитаны значения основных показателей:

– индекс предпринимательства (GEI),

– индекс женского предпринимательства (WEI),

– инфляция и доля женщин в рабочей силе (Female LFPR).

Распределения этих индексов позволяют судить о широте и неоднородности предпринимательских экосистем. Так, средний GEI составляет 47.2 при стандартном отклонении 16.2, а разброс WEI варьируется от 25 до 75, что указывает на значительные межстрановые различия (см. рисунок 1).

	Entrepreneurship Index	Women Entrepreneurship Index	Inflation rate
count	51.00	51.00	51.00
mean	47.24	47.84	2.59
std	16.19	14.27	5.38
min	24.80	25.30	-2.25
25%	31.90	36.35	-0.50
50%	42.70	44.50	0.60
75%	65.40	59.15	3.60
max	77.60	74.80	26.50

Рисунок 1 – Описательная статистика датасета WEI

Гистограммы GEI и WEI (см. рисунок 2 и 3) наглядно демонстрируют бимодальность:

– одна группа стран сосредоточена в диапазоне 25–45,

– другая в диапазоне 60–75.

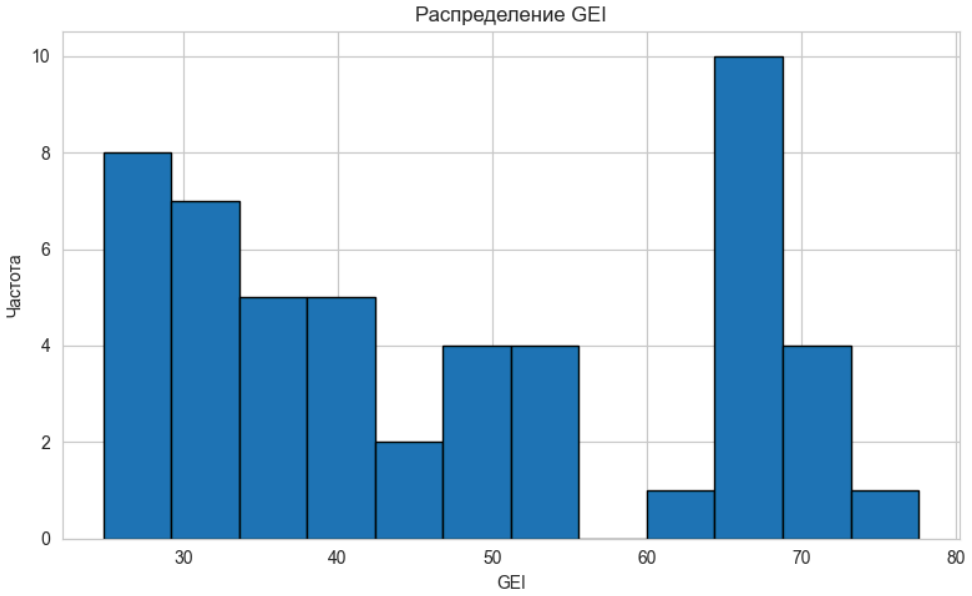

Рисунок 2 – Гистограмма распределения GEI

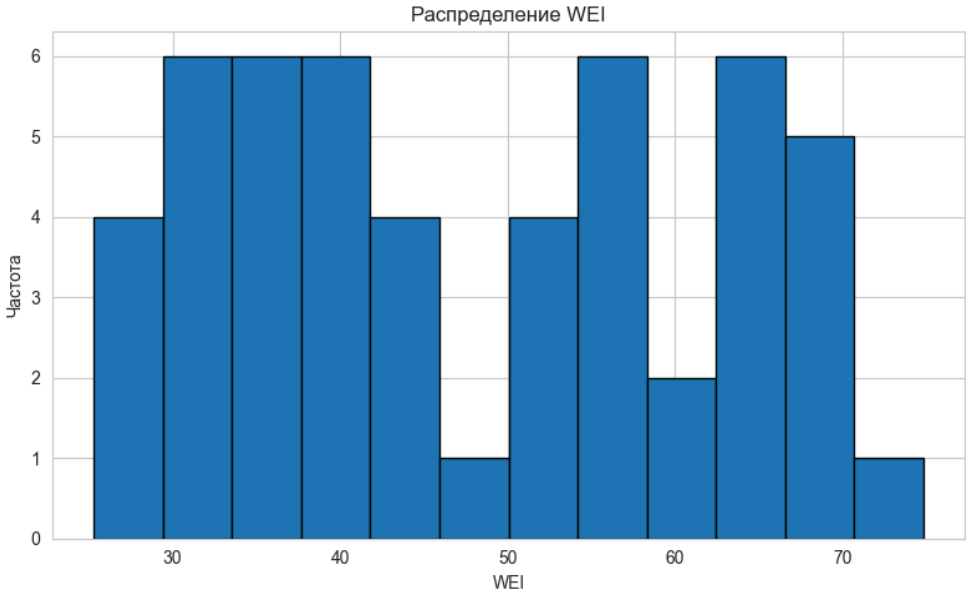

Рисунок 3 – Гистограмма распределения GEI

Такая форма распределения указывает на существование разрывов между кластерами стран, группы с высокоразвитой предпринимательской средой и развивающиеся рынки, где предпринимательство менее распространено.

На гистограмме WEI также видна выраженная бимодальность, однако с небольшим смещением вниз, в большинстве стран индекс женского предпринимательства заметно ниже общего GEI, что иллюстрирует наличие гендерного разрыва в доступе к бизнесу.

Матрица корреляций (см. рисунок 4) позволила выявить степень линейной связи между основными переменными. Самая высокая корреляция наблюдается между GEI и WEI где r ≈ 0.91, что указывает на тесную связь, там, где предпринимательская среда развита в целом, обычно выше и женская активность. Корреляция GEI и Female LFPR также положительна r ≈ 0.33, что подтверждает важность массового участия женщин для развития предпринимательства. Отрицательная корреляция GEI с инфляцией r ≈ –0.4 подчеркивает значимость макроэкономической стабильности, в странах с высокой инфляцией предпринимательская активность заметно ниже. Аналогичные зависимости наблюдаются и для WEI, он положительно коррелирует с LFPR r ≈ 0.44 и отрицательно с инфляцией r ≈ –0.46, что согласуется с гипотезой стабильной среды.

	Entrepreneurship Index
Entrepreneurship Index	1.00
Women Entrepreneurship Index	0.91
Inflation rate	-0.40
Female Labor Force Participation Rate	0.33

	Women Entrepreneurship Index
Entrepreneurship Index	0.91
Women Entrepreneurship Index	1.00
Inflation rate	-0.46
Female Labor Force Participation Rate	0.44

	Inflation rate
Entrepreneurship Index	-0.40
Women Entrepreneurship Index	-0.46
Inflation rate	1.00
Female Labor Force Participation Rate	-0.14

	Female Labor Force Participation Rate
Entrepreneurship Index	0.33
Women Entrepreneurship Index	0.44
Inflation rate	-0.14
Female Labor Force Participation Rate	1.00

Рисунок 4 – Корреляционная матрица датасета WEI

Диаграммы рассеяния между GEI и WEI (см. рисунок 5), GEI и инфляцией (см. рисунок 6), а также GEI и Female LFPR (см. рисунок 7) позволили визуально зафиксировать не только общий характер связи, но и отдельные выбросы и аномалии.

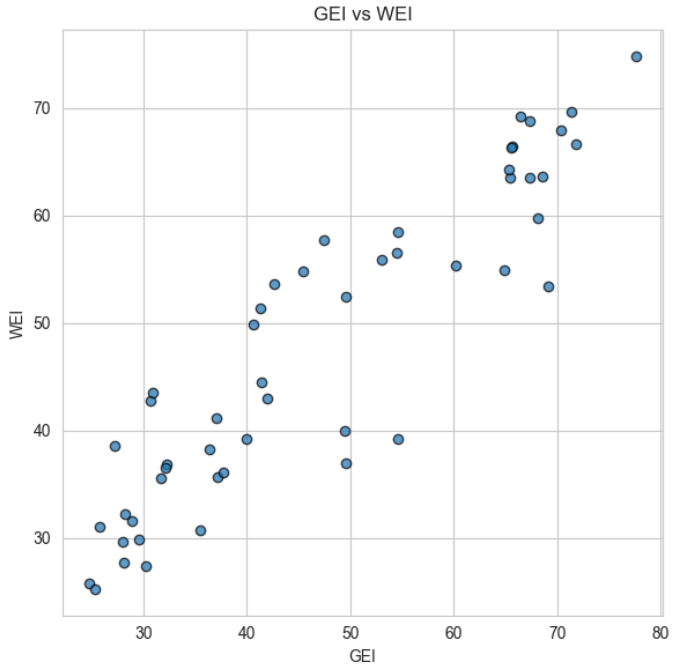

Рисунок 5 – Диаграмма рассеяния GEI и WEI

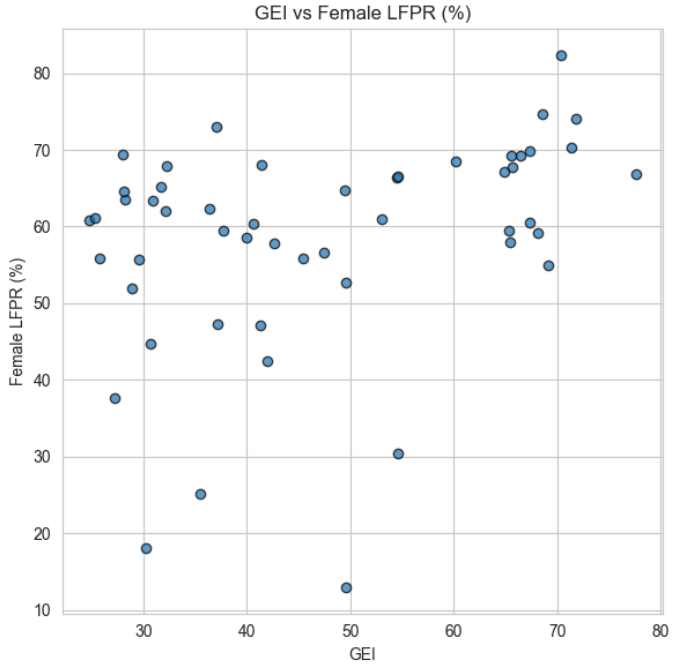

Рисунок 6 – Диаграмма рассеяния GEI и инфляцией

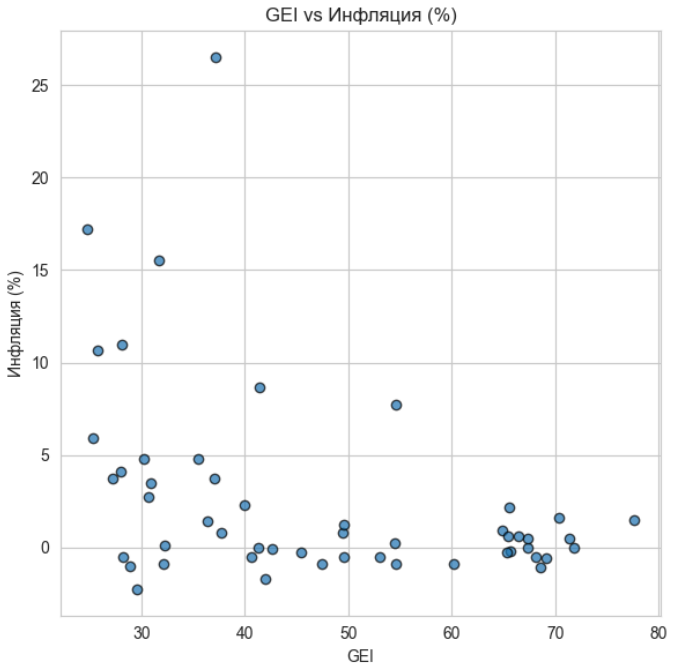

Рисунок 7 – Диаграмма рассеяния GEI и Female LFPR

На графике “GEI и WEI” отчетливо видна линейная зависимость с небольшим числом стран, например, Япония и Индия, где WEI существенно ниже, чем следовало бы ожидать при данном уровне GEI, что говорит о культурных барьерах. Диаграмма “GEI и инфляция” показывает, что страны с низкой или отрицательной инфляцией могут иметь как средние, так и высокие значения GEI, тогда как высокие значения инфляции практически всегда сочетаются с низким GEI. Диаграмма “GEI и Female LFPR” подтверждает гипотезу: страны с долей женщин в рабочей силе выше 60% чаще всего имеют GEI выше 60, что характерно для развитых стран.

Бимодальность в распределении GEI и WEI свидетельствует о том, что страны не распределены равномерно по уровню предпринимательства.

Взаимосвязь между GEI и WEI подчеркивает, что меры по поддержке женского предпринимательства влияют на общий уровень деловой активности.

Высокая инфляция ограничивает предпринимательство, особенно на развивающихся рынках.

Критическим условием высокого GEI выступает массовое включение женщин в экономику, что подтверждается сильной положительной связью GEI и Female LFPR.

5 Построение математических моделей

5.1 Факторный анализ

Был применен факторный анализ с целью выделения скрытых переменных, определяющих основные направления вариации наблюдаемых показателей предпринимательской активности, включая GEI, WEI, инфляцию и долю женщин в рабочей силе. Такой подход позволяет сузить сложное многообразие социальных и экономических характеристик до нескольких интегральных факторов.

Перед анализом все переменные были стандартизированы для обеспечения сопоставимости. Основной критерий Кайзера, eigenvalue > 1 показал наличие одного доминирующего фактора, его собственное значение составило value ≈ 2.43, что объясняет около 61% общей дисперсии переменных (см. рисунок 8).

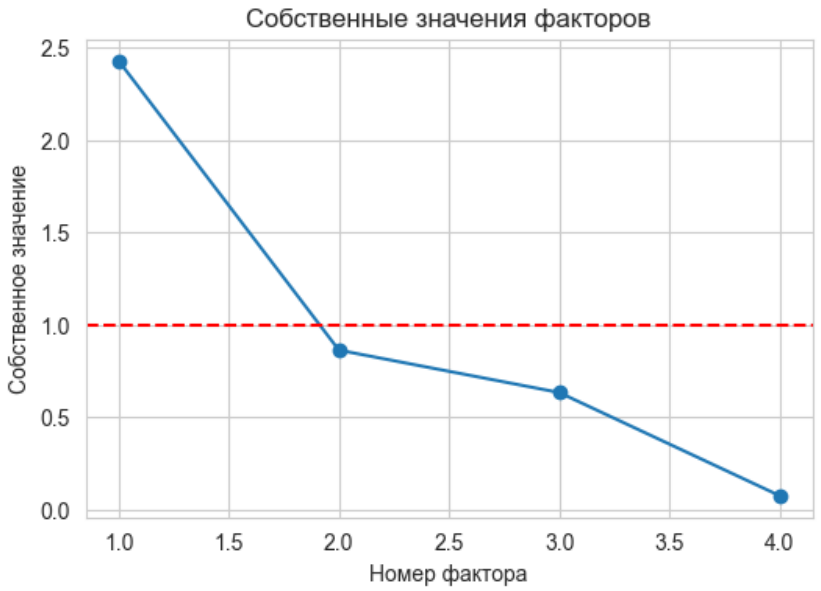

Рисунок 8 – Диаграмма Кайзера

Факторные нагрузки свидетельствуют, что в этот фактор с высокими коэффициентами входят GEI, WEI и Female LFPR положительно, а с отрицательным знаком, что указывает на социокультурно-экономическое развитие как ключевой интегральный фактор (см. рисунок 9).

	1
Entrepreneurship Index	-0.896029
Women Entrepreneurship Index	-1.011144
Inflation rate	0.438626
Female Labor Force Participation Rate	-0.399985

Рисунок 9 – Нагрузки факторов

Страны с высоким значением этого фактора имеют развитую предпринимательскую экосистему, это сочетание инноваций, инклюзивности, особенно женской, и макроэкономической стабильности. Факторный анализ также выявил наличие второстепенного фактора, отражающего специфически гендерные или макроэкономические различия, однако его вклад в общую вариацию незначителен.

На практике, интегральный фактор указывает на необходимость комплексных мер, нельзя добиться прогресса только в одном измерении, например, только увеличить женскую занятость, не развивая прочие компоненты.

5.2 Кластерный анализ

Кластерный анализ был проведен для выявления однородных групп стран по основным характеристикам. Цель – сегментировать страны для последующего формирования адресных политик поддержки предпринимательства.

Использование elbow-метода (см. рисунок 10) позволило определить оптимальное число кластеров, два кластера с максимальным различием, что подтверждается изломом инерции при k = 2.

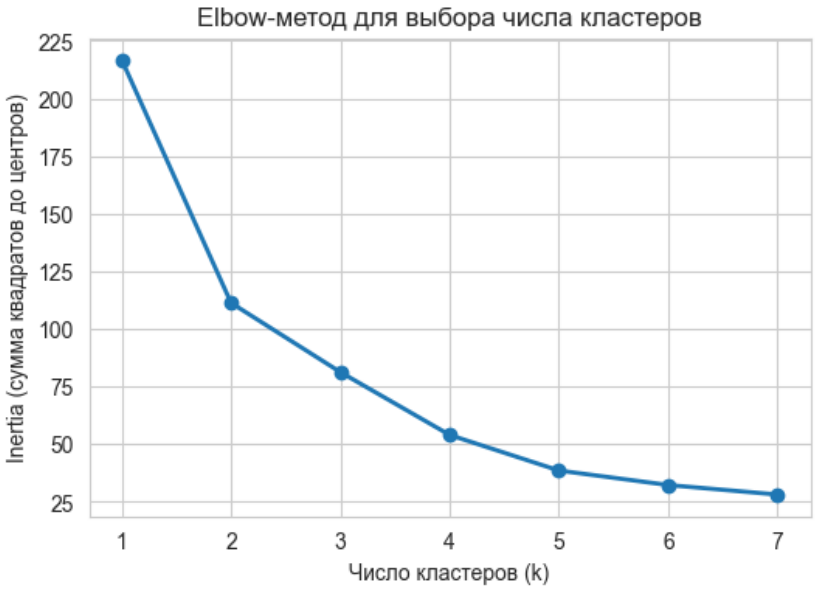

Рисунок 10 – Диаграмма elbow-метода

Кластер 1 имеет страны с высокими показателями GEI и WEI, минимальной инфляцией и значительной долей женщин в рабочей силе. Это в основном развитые экономики, как пример, страны Северной/Западной Европы и США.

Кластер 2 определяет страны с низкими индексами GEI и WEI, высокой инфляцией и ограниченным женским участием в экономике, такие как, Индия и Турция.

Кластеризация отражает различия как по экономическому, так и по социокультурному профилю. Она дополнительно визуализирована на диаграмме рассеивания GEI и WEI (см. рисунок 11).

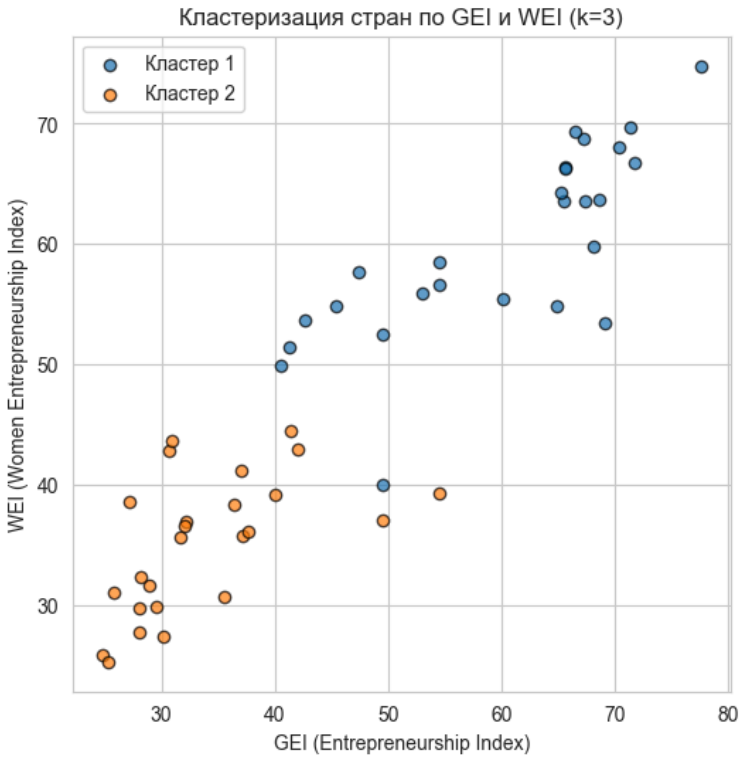

Рисунок 11 – Диаграмма рассеивания GEI и WEI с кластеризацией

5.3 Классификационные модели

На микроуровне для датасета студентов строились модели классификации для прогноза склонности студентов к предпринимательству на основе их личностных и социокультурных характеристик. Это позволяет не только выделить ключевые предикторы на уровне личности, но и протестировать, насколько современные методы могут автоматизировать отбор потенциальных предпринимателей для участия, например, в образовательных программах.

Использовались следующие модели:

– логистическая регрессия,

– дерево решений,

– random forest.

Лучшие результаты показал Random Forest, точность на тесте – 54%, AUC – 0.51 (см. рисунок 12).

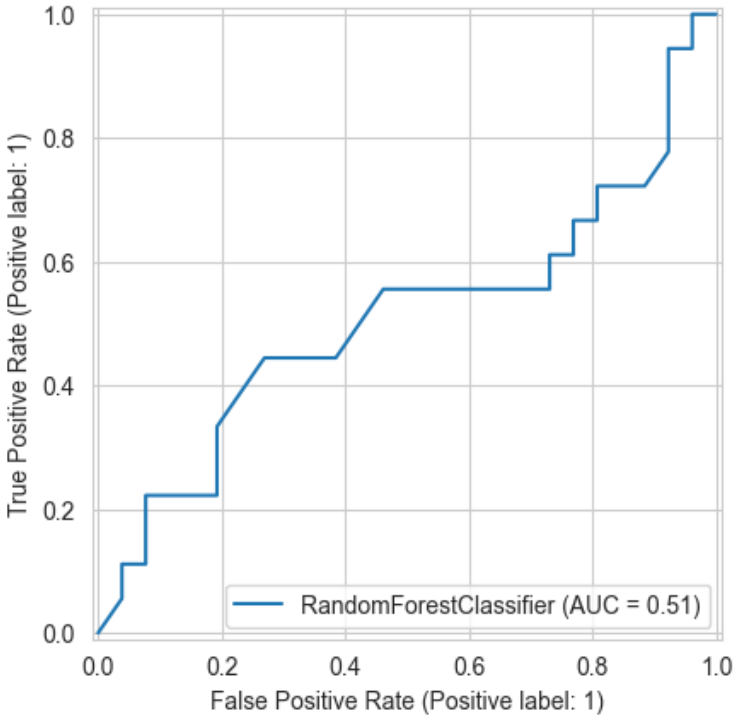

Рисунок 12 – Диаграмма с ROC кривой

Значимость признаков (см. рисунок 13) выявила ключевые факторы:

– возраст,

– хорошее здоровье,

– инициативность,

– самоуверенность.

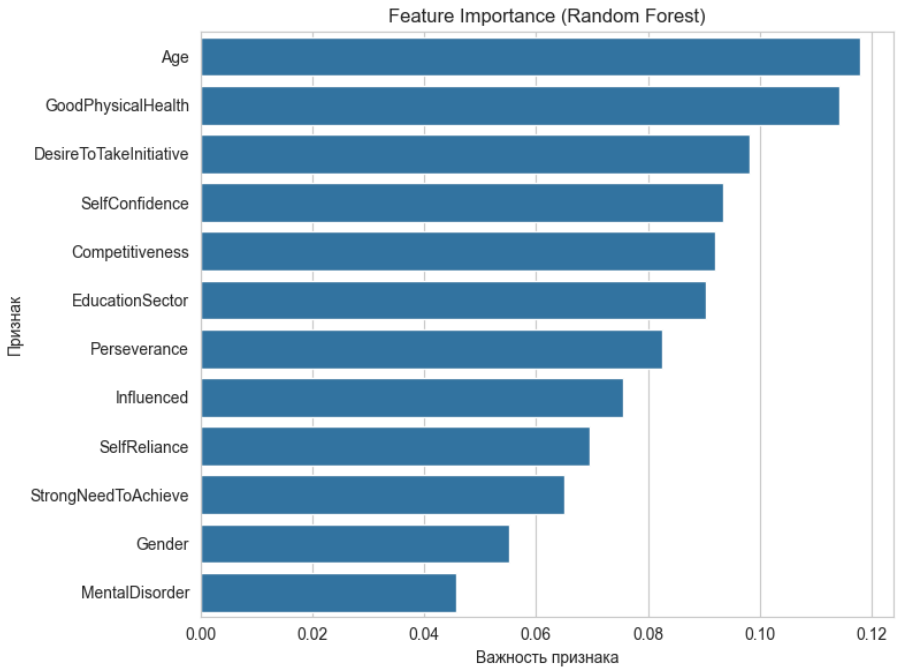

Рисунок 13 – Столбчатая диаграмма значимости признаков RF

Пол и наличие ментальных расстройств не оказали существенного влияния.

Небольшой объем данных дал низкую точность. Для повышения прогностической силы моделей необходим сбор более информативных данных.

5.4 Регрессионные и сценарные модели

В рамках регрессионного анализа с использованием множественной линейной регрессии и Random Forest Regressor оценивалась объясняющая сила различных факторов в предсказании GEI. Применялись как стандартные статистические методы OLS, так и машинное обучение для учета возможных нелинейных связей.

Модель OLS показала, что WEI является наиболее значимым фактором, рост WEI на 1 пункт приводит к увеличению GEI на 1.07 (см. рисунок 14).

	coef	std err	t	P> |t|	[0.025	0.975]
const	-2.3207	5.513	-0.421	0.676	-13.502	8.861
Women Entrepreneurship Index	1.0724	0.092	11.644	0.000	0.886	1.259
Inflation rate	0.0969	0.210	0.461	0.648	-0.329	0.523
Female Labor Force Participation Rate	-0.0285	0.091	-0.312	0.757	-0.214	0.157

Рисунок 14 – значения модели OLS

Female LFPR и инфляция статистически значимого вклада не внесли, хотя направления их влияния соответствуют теоретическим ожиданиям.

В модели Random Forest важность WEI подтверждается, на нее приходится более 90% вклада (см. рисунок 15).

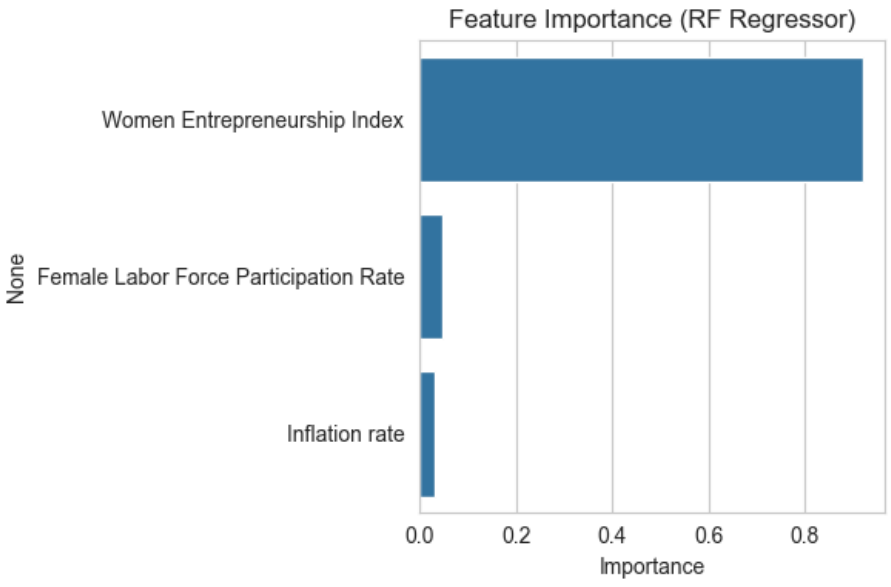

Рисунок 15 – Столбчатая диаграмма значимости признаков RFR

Сценарный анализ показал, что при увеличении WEI, даже при неизменной инфляции и LFPR, GEI растет практически линейно (см. рисунок 16).

	Women Entrepreneurship Index	Inflation rate	\
0	20	0	
1	40	2	
2	60	5	
3	80	10	

	Female Labor Force Participation Rate	GEI_pred_linear	GEI_pred_rf
0	58.481765	17.462295	28.2235
1	58.481765	39.104134	40.7615
2	58.481765	60.842854	56.3055
3	58.481765	82.775334	72.3385

Рисунок 16 – Значения сценарного анализа

ЗАКЛЮЧЕНИЕ

Проведенный анализ показал, что предпринимательская активность тесно связана с социокультурными факторами, в первую очередь с уровнем гендерной инклюзии и макроэкономической стабильностью. Индекс женского предпринимательства оказался главным фактором, влияющим на общий уровень предпринимательства в стране.

Использование математических моделей позволило выявить профили стран и определить, какие параметры требуют поддержки в разных группах государств. В развитых странах приоритетом становятся равные возможности, а в развивающихся, улучшение делового климата с повышением женской занятости.

Приложение А

Репозиторий с исходным кодом: [ссылка]