Санкт-Петербургский государственный университет

ГЛУШКОВ Егор Александрович

Выпускная квалификационная работа

Определение кода Голланда по результатам психометрических тестов личности на основе методов машинного обучения в условиях неполноты информации

Уровень образования: магистратура

Направление *02.04.03 «Математическое обеспечение и администрирование информационных систем»*

Основная образовательная программа *ВМ.5665.2023 «Математическое обеспечение и администрирование информационных систем»*

Научный руководитель:
доцент кафедры информатики, к. т. н., Абрамов М. В.

Консультант:
ст. преподаватель кафедры информатики, Столярова В. Ф.

Рецензент:
ст. научный сотрудник, СПб ФИЦ РАН, к. т. н., Захаров В. В.

Санкт-Петербург
2025

Saint Petersburg State University

Egor Glushkov

Master's Thesis

Determination of the Holland Code based on the results of psychometric personality tests using machine learning methods in the presence of incomplete information

Education level: master

Speciality *02.04.03 "Software and Administration of Information Systems"*

Programme *BM.5665.2023 "Software and Administration of Information Systems"*

Scientific supervisor:
C.Sc., docent, Department of CS, M. V. Abramov

Consultant:
senior lecturer, Department of CS, V. F. Stoliarova

Reviewer:
C.Sc., senior researcher, SPb FRC RAS, V. V. Zakharov

Saint Petersburg
2025

Оглавление

Введение

Многие аспекты успешности человека обусловлены корректным определением карьерного пути, который соответствует его личностным предпочтениям [1]. От того, насколько успешно человек определил свою социально-профессиональную направленность, зависит удовлетворенность человека своей работой [2–4]. Карьерному самоопределению уделяются значительные ресурсы, в том числе со стороны государства [5], чьи усилия направляются в том числе на профессиональное образование людей, которые по окончании обучения не работают по своей специальности вследствие отсутствия интереса к выбранной профессии. С определением своего будущего с профессиональной точки зрения сталкивается любой выпускник школы и вуза, безработный, работник, не удовлетворенный текущей работой или уже находящийся в процессе её смены, — для всех этих категорий людей встает вопрос об определении своих профориентационных предпочтений. Золотым стандартом является глубинное интервью с экспертом, который поможет выявить сильные и слабые стороны личности. Однако этот подход является ресурсозатратным, и для автоматизации процесса профориентации используются дистанционные способы карьерного консультирования [6, 7], в том числе профориентационные тесты, доступные для прохождения онлайн и не требующие большого количества времени для прохождения [8].

Одним из инструментов для определения профессиональных интересов уже более полувека является модель Дж. Голланда RIASEC [9], которая предполагает, что профессиональные предпочтения являются отражением характера человека, его базовых черт. Так, вводятся шесть типов социально-профессиональной направленности личности (кодов Голланда): реалистический (Realistic, R), исследовательский (Investigative, I), артистический (Artistic, A), социальный (Social, S), предприимчивый (Enterprising, E) или традиционный (Conventional, C). Коды могут определяться при помощи тестирования, в котором в результате попарного сравнения профессий численно оцениваются указанные шесть типов направленности личности [10]. Для сравнения профессиональных профилей личности (кодов Голланда) между собой ис-

пользуется C-индекс как мера сходства (конгруэнтности). Существует множество вариаций данного теста [11], результаты которых коррелированы, однако не определяют друг друга однозначно. Кроме того, такие тесты часто не учитывают быстро изменяющуюся конъюнктуру рынка профессий, а также культурные и социо-экономические различия респондентов [11–13]. Возникает актуальная задача определения кода Голланда по альтернативным данным.

В настоящее время существуют исследования, показывающие взаимосвязь кода Голланда индивида с его социально-демографическими признаками [14], цифровым следом в социальных сетях (его сообщения, посты, фото; в исследованиях могли быть использованы иные психометрические тесты) [15–21]. Однако основное внимание исследователей направлено на изучение взаимосвязи факторов модели Голланда с факторами других психометрических тестов [22–28]. Для анализа взаимосвязей различных опросных инструментов исследователи используют методы регрессионного анализа и статистические тесты [16, 17, 23], моделирование структурными уравнениями (*SEM*) [11, 28] и модели машинного обучения [14, 18–21, 25, 29, 30].

Несмотря на наличие работ, в которых предсказывается результат одного психометрического теста на основе другого теста или на основе некоторых признаков личности (например, комментариев, постов и фото пользователей социальной сети), до сих пор нет инструментов, позволяющих по результатам одного или комбинации сразу нескольких популярных психометрических тестов («Большая пятёрка», Кеттелла, Айзенка, Леонгарда, Шварца) предсказывать код Голланда. Востребован инструмент, в котором результаты тестов могли бы быть предоставлены частично, который бы в меньшей степени зависел от изменяющейся конъюнктуры рынка профессий, от культурных и экономических различий респондентов, то есть позволял определять профессиональный профиль личности в условиях подобной неполноты информации. В результате, пользователь мог бы получить информацию о своих профориентационных предпочтениях без прохождения теста Голланда. Кроме того, даже при прохождении последнего подобный инструмент может использоваться для уточнения результатов теста Голланда и установ-

ления его непротиворечивости в соответствии с результатами других исследований.

Данная работа призвана закрыть этот пробел: с помощью современных методов машинного обучения на основе собранных данных по результатам прохождения одного или нескольких указанных психометрических тестов личности устанавливаются взаимосвязи между результатами тестов и предсказывается код Голланда, соответствующий профессиональным предпочтениям личности. Предсказание кода Голланда может быть решено такими методами машинного обучения, как многоцелевая регрессия, классификация или ранжирование [31]; для улучшения качества прогноза базовые модели могут объединяться в ансамбли [32, 33]. Возможность предсказания по одному или нескольким тестам влечет необходимость восстановления результатов тестов, которые не были пройдены.

Новизна результатов исследования состоит в создании нового программного комплекса, обеспечивающего автоматизацию процесса профориентации на основе предсказания кода Голланда. Теоретическая значимость заключается в использовании уникальной комбинации различных психометрических тестов при разработке новых моделей машинного обучения для определения взаимосвязи тестов и кода Голланда. Практическая значимость — разработка прототипа программного модуля автоматизации оценки профессиональной направленности по психологическому профилю личности.

1. Постановка задачи

Целью работы является автоматизация процесса профориентации посредством разработки инструмента для предсказания кода Голланда по неполным результатам психометрических тестов с использованием методов машинного обучения.

Для выполнения цели были поставлены следующие задачи:

1. Разработать математическое обеспечение для модуля восстановления пропусков психометрических тестов.
2. Реализовать различные подходы к определению кода Голланда: многоцелевая регрессия, классификация, ранжирование.
3. Разработать модуль формирования взвешенного ансамбля моделей для объединения прогнозов базовых алгоритмов.
4. Провести сравнительный анализ подходов и методов определения кода Голланда на основе C-индекса.
5. Создать прототип инструмента для определения профориентационных предпочтений личности.

2. Обзор

В данном разделе представлен обзор модели Голланда и дано краткое описание психометрических тестов личности. Проведен обзор предметной области: рассмотрены и проанализированы статьи, предлагающие различные подходы к решению задачи предсказания значений факторов психометрических тестов и нахождения взаимосвязей между факторами.

2.1. Модель Голланда и психометрические тесты личности

2.1.1. Модель Голланда RIASEC

Одним из основных инструментов оценки профессиональных интересов служит модель Голланда, также известная как модель *RIASEC*. Данная методика была разработана Джоном Льюисом Голландом в конце 1950-х годов [9], после чего им неоднократно дорабатывалась и развивалась. В своей статье «Теория профессионального выбора» исследователь сопоставляет различным типам личности профессиональные роды деятельности. Согласно Голланду, личности выбирают и преуспевают в той профессиональной среде, которая подходит их характеру, является отражением их базовых черт, при этом профессиональная карьерная среда классифицируется по тем типам личностей, которые в этой среде успешны. Таким образом, для определения профессиональных предпочтений достаточно определить социально-профессиональный тип личности.

В своих более поздних работах ученый выделяет следующие шесть типов личностей:

- реалистический (*Realistic*, R);
- исследовательский (*Investigative*, I);
- артистический (*Artistic*, A);
- социальный (*Social*, S);
- предприимчивый (*Enterprising*, E);
- традиционный (*Conventional*, C).

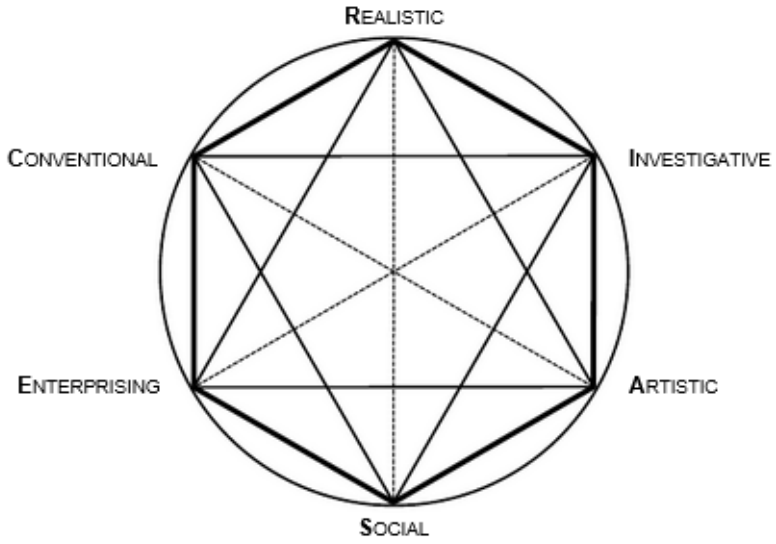

Рисунок 1. Правильный шестиугольник, в вершинах которого расположены типы личности (коды) согласно Голланду[1]

При этом определяется не единственный тип личности: оценивается принадлежность человека к каждому из типов, которые затем выстраиваются в порядке убывания их выраженности. Результат записывается в виде кода по первым буквам типов, что и дало название модели. Модель Голланда также можно представить как правильный шестиугольник с кодами в вершинах (рисунок 1).

Существуют различные методики определения кода Голланда, и в российской адаптации методики Г. В. Резапкиной [10] предлагается опросный инструмент с 42 парами профессий: из каждой пары опрашиваемый должен выбрать одну профессию, которая его более привлекает. Каждой из выбранных профессий сопоставляется один из шести типов, соответствие профессий и типов было произведено Голландом в своих работах и дополнено другими исследователями. Для каждого из типов подсчитывается количество упоминаний, и шесть типов выстраиваются в порядке убывания подсчитанных баллов. Например, опрашиваемый 11 раз в парах выбрал профессии, соответствующие исследовательскому типу (I), 10 — традиционному (C), 9, 7, 5 и 0 предприимчивому (E), реалистическому (R), социальному (S) и артистическому (A) типам соответственно. Итоговым кодом для опрашиваемого будет «ICERSA»,

[1]Источник: Hexagonal arrangement of the RIASEC dimensions. URL: https://www.researchgate.net/figure/Hexagonal-arrangement-of-the-RIASEC-dimensions_fig1_356677137

который можно ограничить тремя первыми буквами («верхняя триада», «ICE»).

Для сравнения кода типа личности и кода его профессиональной среды Голланд вводит понятие конгруэнтности (согласованности). Среди мер согласованности можно выделить C-индекс для трёхбуквенных кодов («верхних триад»):

$$C = 3\,(X_1, Y_1) + 2\,(X_2, Y_2) + 1\,(X_3, Y_3),$$

где $\{X_i\}$ и $\{Y_i\}$ — первые три позиции кодов Голланда, их позиции в замкнутой цепочке (шестиугольнике) *R-I-A-S-E-C*:

$$(X_i, Y_i) = \begin{cases} 3, & \text{если } X_i = Y_i, \\ 2, & \text{если } X_i \text{ и } Y_i \text{ – соседние позиции}, \\ 1, & \text{если } X_i \text{ и } Y_i \text{ – позиции через один код}, \\ 0, & \text{если } X_i \text{ и } Y_i \text{ – противоположны}. \end{cases}$$

Бóльшие значения меры согласованности говорят о большем сходстве двух профессиональных профилей (кодов Голланда).

2.1.2. Психометрические тесты личности

Помимо модели Голланда есть и другие способы определения профориентационных предпочтений. Одним из таких способов являются психометрические опросные инструменты. Их основная цель — отразить некоторые черты личности человека в удобном числовом формате. Они помогают выявить ключевые черты характера, темперамент, ценности и поведенческие особенности.

Среди психометрических тестов личности можно выделить следующие (в скобках указано количество факторов):

1. Опросник Леонгарда-Шмишека (10).
2. Личностный опросник Айзенка (4).
3. 16-факторный опросник Кеттелла (16).
4. Пятифакторный опросник личности («Большая пятёрка»; 5).
5. Ценностный опросник Шварца (20).

Опросник Леонгарда-Шмишека разработан Г. Шмишеком на основе теории акцентуаций личности К. Леонгарда [34] (адаптация на русский язык — [35]) и представляет собой опросник из 88 вопросов с ответами «да/нет». Вопросы сгруппированы по 10 шкалам, по каждой из которых подсчитывается балл, отражающий выраженность соответствующей акцентуации. Методика позволяет составить индивидуальный психологический портрет, выявить усиленные черты характера, влияющие на поведение, эмоциональные реакции и адаптацию. Применяется в образовательных программах, при подборе персонала и профориентации.

Личностный опросник Айзенка (EPQ-R) разработан Г. Айзенком и коллегами в 1985 г. [36] (русская адаптация Суходольского [37]) и направлен на измерение базовых личностных параметров: экстраверсии, нейротизма, психотизма и оценки искренности ответов. Опросник включает 100 вопросов с ответами «да/нет», по результатам которых формируются соответствующие шкалы, позволяющие оценить стиль взаимодействия человека с окружающим миром, его стрессоустойчивость и темперамент. Методика широко применяется в клинической психологии, образовательных программах, профориентации и научных исследованиях.

16-факторный опросник Кеттелла (16PF) [38] (адаптация [39]) предназначен для всесторонней диагностики личности через измерение 16 основных черт характера методом факторного анализа. Тест состоит из 187 вопросов с ответами «утвердительно», «отрицательно» и «нейтрально», что позволяет получить детальный личностный профиль. По итогам подсчитываются баллы по каждой из 16 шкал (от 0 до 26) и дополнительным шкалам, отражающим вторичные личностные особенности. Инструмент широко применяется при кадровом отборе, психологическом консультировании и научных исследованиях для глубокого понимания индивидуальных различий. Именно 16PF послужил основой для появления пятифакторных тестов («Большая пятёрка»).

Пятифакторный опросник личности («Большая пятёрка», 5PFQ) [40] (русская адаптация Хромова [41]) основан на модели факторного анализа *NEO-PI-R* [42] и представляет собой диспозиционную методику для оценки пяти базовых черт личности: экстраверсия,

привязанность, самоконтроль, эмоциональная устойчивость, экспрессивность. Тест включает 75 утверждений, которые оцениваются от –2 до 2, что позволяет получить итоговые шкалы, отражающие степень выраженности каждой черты, а также множество вторичных факторов. Инструмент широко применяется в профориентации, подборе персонала и научных исследованиях.

Ценностный опросник Шварца (SVS) [43] (русская адаптация Карандашева [44]) предназначен для количественной оценки 10 универсальных ценностей, выявленных методом исследования жизненных приоритетов. Методика использует парное сравнение ценностей и фиксирует две оценки: нормативный идеал (какие ценности человек считает важными в идеале) и индивидуальный приоритет (что важно для него лично). Опросник активно применяется в социологических и психологических исследованиях для анализа культурных различий и индивидуальных мотивационных профилей.

2.2. Методы машинного обучения в вычислительной психологии личности

Применению методов машинного обучения в вычислительной психологии личности (*personality computing*) в целом и нахождению взаимосвязей результатов различных психометрических тестов между собой в частности посвящено множество научных работ. Для решения задачи установления взаимосвязи между факторами психологических тестов, а также внешними признаками используются следующие методы:

- Статистические: регрессионный, корреляционный, факторный анализ; структурное моделирование (*SEM*).
- Динамические: модели, описывающие изменения во времени (дифференциальные уравнения).
- Сетевые: анализ взаимосвязей между переменными как узлов в сети.
- Машинное обучение: классификация, кластеризация, регрессия, ранжирование.

Так, всё чаще встречается применение методов машинного обучения

в вычислительной психологии. Задачи могут быть различны.

- Предсказание кода Голланда на основе социально-демографических признаков [14]. Авторы предлагают различные подходы для решения этой задачи: с тех пор как код Голланда может быть представлен и как последовательно идущие три или шесть кодов, и как целочисленные значения кодов, то и задача может быть поставлена следующим образом: регрессия с множественными выходами (*многоцелевая*; *multioutput regression*), классификация с несколькими метками (*multilabel classification*), классификация с множественными выходами (*multioutput classification*). Авторы отмечают: в случае последовательного предсказания для многоцелевой регрессии порядок предсказания выходов важен. В качестве меры сходства авторы используют меру конгруэнтности — C-индекс. Наилучшие результаты показал градиентный бустинг: C-index = 10.95 при решении задачи регрессии и C-index = 11.08 при решении задачи классификации.

- Оценка профессионального выбора [30]. Пользователю по результатам прохождения теста Голланда предъявлялся список профессий, которым прежде уже был сопоставлен свой код Голланда; требовалось найти наиболее подходящие профессии. Наилучших результатов удалось достичь с помощью комбинации традиционных методов и ансамбля методов машинного обучения. В качестве традиционных методов использовалось сравнение значений мер конгруэнтности (простое совпадение главного фактора кода Голланда, оценка профилей — числовых значений кода Голланда — с помощью таких метрик, как коэффициент корреляции Пирсона и Евклидово расстояние). В ансамбль методов машинного обучения вошли следующие модели: многослойный перцептрон (нейронная сеть), метод k-ближайших соседей, регуляризованная регрессия, случайный лес.

- В статье [25] применяется логистическая регрессия для предсказания (классификации), какой путь выберут учащиеся: академи-

ческий или профессиональный; предикторами служили значения факторов тестов Голланда и «Большой пятерки».

- Расширение списка профессий, поставленных в соответствие кодам Голланда, путем создания платформы для автоматизации профилирования вакансий [29]. Предсказание кодов Голланда решается как задача ранжирования с метрикой NDCG (*Normalized Discounted Cumulative Gain*).

- Предсказание значений шкал теста «Большой пятерки» пользователей социальных сетей по следующим признакам: их посты, комментарии, репосты и численные характеристики аккаунта пользователя. Решалась задача бинарной классификации (шкалы теста были представлены бинарными путем сравнения с пороговым значением) с использованием моделей случайного леса и метода опорных векторов [19]. Подобная задача в работе [21] решалась с помощью многослойного перцептрона. По схожим признакам (в т. ч. по указанной в профиле пользователя информации) для оценки темперамента (тест Айзенка EPQ) в статье [18] использовались модели CatBoost и случайный лес. В работе [20] предсказание результатов тестов «Большой пятерки», Шварца и других по извлекаемым из профилей в социальной сети численным признакам (число друзей, постов, подписок, длина поля с личным описанием, длина постов и др.) осуществляется с помощью модели XGBoost (*eXtreme Gradient Boosting*).

2.3. Выводы

Для предсказания кода Голланда могут быть использованы различные идеи, рассмотренные в данном обзоре. Например, задачу можно сформулировать как регрессию/классификацию с множественными выходами (*multioutput*), классификацию с несколькими метками (*multilabel*). Для многоцелевой регрессии могут быть использованы различные метрики: не только усредненные MSE или RMSE, часто применяемые в таких задачах, но и специальные меры конгруэнтности

(*сходства*, C-индекс), косинусное расстояние, коэффициент корреляции. Приведены различные методы машинного обучения, в том числе и те, с помощью которых были достигнуты наилучшие результаты: в первую очередь, это градиентный бустинг (CatBoost, XGBoost) и случайный лес. В соответствии с результатами работы [14] приемлемым может считаться следующий результат предсказания значений кода Голланда: C-index ≥ 11.

Несмотря на разнообразие работ, наличие среди них тех, где по результатам одних психометрических тестов предсказываются другие, до сих пор нет инструментов, позволяющих по результатам сразу нескольких популярных психометрических тестов — «Большой пятерки», Кеттелла, Айзенка, Леонгарда, Шварца — предсказать код Голланда. Таким образом, пользователь мог бы получить информацию о своих профориентационных предпочтениях без прохождения теста Голланда. Кроме того, даже при прохождении последнего подобный инструмент мог бы уточнять результаты теста Голланда, проверять его непротиворечивость в соответствии с результатами других тестов.

3. Модели и методы определения кода Голланда по данным психометрических тестов

В данном разделе описываются различные подходы к предсказанию кода Голланда на полных данных (без пропусков) и на данных с пропусками, требующими восстановления. В конце раздела приводится итоговая структура решения задачи по предсказанию кода Голланда.

3.1. Определение кода Голланда на полных данных: регрессия, классификация, ранжирование

Данные психометрических тестов представляют собой набор признаков, принимающих целочисленные значения. Пример данных психометрических тестов приведен в таблице 1. Определение кода Голланда представляет собой предсказание шести чисел, отражающих степень выраженности типов личности (кодов Голланда). В качестве альтернативы набору из шести чисел предсказанием кода может служить ответ как в виде однобуквенного значения, так и набора из трех букв, «верхней триады» — тройки наиболее выраженных кодов. В случае, если в таком наборе задан порядок, то речь может идти о предсказании рангов кодов (от менее выраженного к наиболее выраженному).

Таблица 1. Пример данных психометрических тестов

id	**Бол. пятёрка**			**...**	**Леонгард**		**Голланд**					
	BF1	BF2	BF3		LN9	LN10	HL1	HL2	HL3	HL4	HL5	HL6
1	39	66	33	...	3	12	8	8	6	8	1	11
2	45	46	73	...	12	6	3	7	7	8	10	7
3	34	41	56	...	18	12	10	10	3	11	7	1
4	49	47	50	...	15	24	6	4	8	6	7	11
5	48	42	53	...	12	6	6	7	8	7	10	4

Таким образом, задача определения кода Голланда по результатам психометрических тестов личности может быть сведена к следующим

задачам:

1) регрессия со множественными выходами (набор чисел);

2) классификация (набор кодов);

3) ранжирование (упорядоченный набор кодов);

4) ансамблевые модели (комбинация базовых моделей).

3.1.1. Регрессия

Предсказание сразу нескольких целевых числовых переменных представляет собой регрессию со множественными выходами (многоцелевая регрессия, *multitarget*). Существует три основных подхода к решению данной задачи [31, 45]:

- Применение моделей, по умолчанию поддерживающих множественные выходы (линейная регрессия, метод k-ближайших соседей, нейронные сети).
- Независимые предсказания каждого из выходов (*multioutput*).
- Предсказания выходов по цепочке: последний предсказанный выход становится частью признакового пространства для предсказания следующего выхода (*regressor chain*, см. рисунок 2).

Оцениваются результаты решения задачи регрессии со множественными выходами с помощью следующих усредненных по ответам метрик на тестовой выборке:

- усредненная среднеквадратичная ошибка (RMSE);
- усредненный C-индекс (мера сходства, см. 2.1.1).

Лучшее качество обеспечивается при минимальных значениях RMSE и при максимальных значениях C-индекса (как мера согласованности). Чем больше C-индекс, тем большее сходство имеют два сравниваемых между собой профиля. Использование C-индекса позволяет сравнивать результаты не только с другими регрессионными задачами,

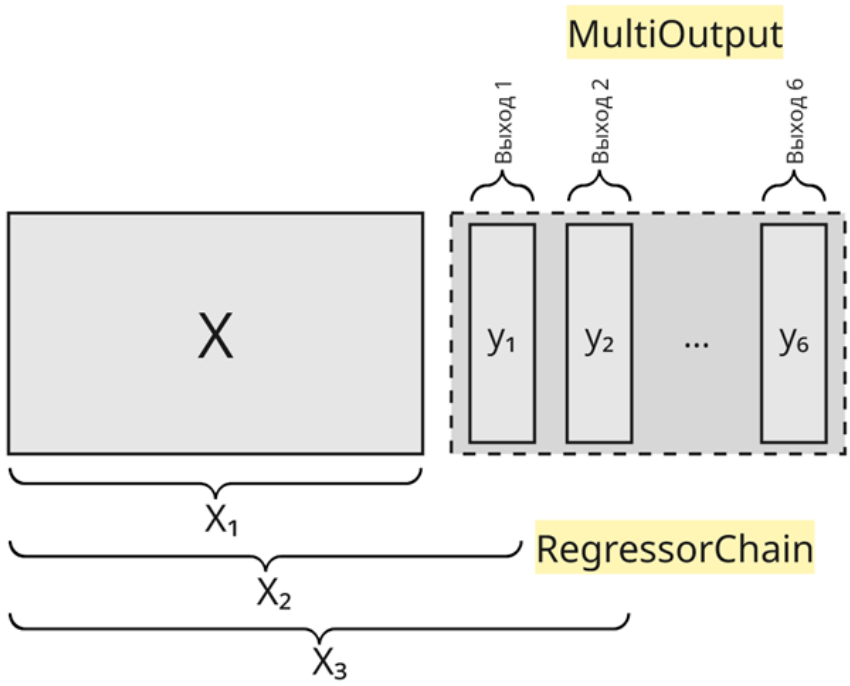

Рисунок 2. Многоцелевая регрессия

но и, например, с задачами классификации. Стоит отметить, что в процессе обучения некоторые модели также позволяют оценить важность признаков: ансамблевые модели на основе решающих деревьев (случайный лес, градиентные бустинги) и линейные регрессоры.

Для нахождения взаимосвязей между факторами модели Голланда и другими психометрическими тестами в качестве базовых были использованы следующие регрессионные модели:

- модели на основе линейной регрессии (простая линейная, пошаговая, L1- и L2-регуляризованные модели регрессии);
- модели на основе ансамблей деревьев с бутстрэппингом (случайный лес Random Forest, ExtraTrees);
- модели градиентного бустинга (XGBoost, LightGBM, CatBoost);
- непараметрические модели (kNN, SVR);
- нейросетевые модели (MLP, *foundation-модель* TabPFN);
- базовая константная модель для сравнительного анализа.

При отсутствии пропусков (то есть при работе с «полными» данными) может быть целесообразным уменьшение размерности с помощью метода главных компонент (англ. *Principal Component Analysis*, PCA) [31]. Некоторые факторы различных тестов могут оценивать одни и те же аспекты личности, среди связей могут наблюдаться корреляции. Уменьшение размерности позволяет агрегировать схожие факторы тестов.

3.1.2. Классификация

Предсказание одной или нескольких категориальных переменных (кодов, классов, меток) представляет собой задачу классификации. Предсказание кода Голланда как решение задачи классификации может быть сведено к следующим подходам [46] (таблица 2):

- Многоклассовая (*multiclass*) классификация.
  Обучается один классификатор на шесть классов. В качестве ответа в порядке убывания вероятностей выбираются три наиболее вероятных кода.

- Многометочная (*multilabel*) классификация.
  Обучаются шесть бинарных классификаторов, позволяющих дать ответ на вопрос, входит ли соответствующий код в «верхнюю триаду». Выбираются три кода с наивысшими предсказанными вероятностями.

- Классификация на полное множество комбинаций (*label powerset*).
  Всего существует 20 различных трехбуквенных комбинаций кода Голланда, в которых не учитывается порядок кодов. На выходе — тройка кодов. По причине отсутствия возможности учесть порядок кодов, для данного подхода невозможно использовать C-индекс, оценка которого основывается именно на порядке кодов.

По аналогии с решением задачи регрессии был применён метод главных компонент. Для оценки решения задачи классификации учитывалась метрика точности Top-K (*Top-K accuracy*), которая показывает, в

Таблица 2. Подходы к классификации

Классификация	**Модель**	**Значение**	**Пример**
Многоклассовая (multiclass)	1 классификатор на 6 классов	3 однобуквенных кода	`R, S, E`
Многометочная (multilabel)	6 бинарных классификаторов	Булевый вектор из 6 элементов с 3 значениями `True`	`[T, F, F, T, T, F]`
Множество комбинаций (label powerset)	1 классификатор на 20 классов	трехбуквенный код, порядок кодов не важен	`RSE`

скольких случаях среди трех наиболее вероятных предсказанных моделью классов есть хотя бы K правильно предсказанных меток. Например, Top-2_{acc} = 0.7 означает, что модель для 70% предсказанных троек кодов угадала как минимум два истинных кода. Для сравнения с результатами других подходов также использовался C-индекс.

Для задачи классификации были взяты те же модели, что и для регрессии, за исключением нейросетевых моделей, а также с добавлением регуляризованных логистических регрессий (вместо линейных регрессий) и наивного байесовского классификатора.

3.1.3. Ранжирование

Представим задачу определения кода Голланда как задачу списочного ранжирования (*listwise learn-to-rank*) [47]. В отличие от поточечного ранжирования, где каждый элемент оценивается по шкале релевантности (что фактически сводится к задаче регрессии), или от попарного ранжирования, в котором моделируется относительный порядок пар элементов (бинарная классификация), списочное ранжирование рассматривает весь список результатов как единый упорядоченный объект.

Списочный подход требует задания двух ключевых компонентов: определения скоринговой функции и выбора функции потерь, оптими-

зирующей качество выдачи списка. В качестве скоринговой функции обычно используется многослойный перцептрон или более сложные архитектуры: глубокая перекрёстная сеть (*Deep & Cross Network*) либо трансформер с механизмом *self-attention*.

Типичными функциями потерь для списочного ранжирования являются NDCG (нормализованный дисконтированный совокупный прирост) и её дифференцируемые аппроксимации (ApproxNDCG, LambdaRank), а также специальные дифференцируемые функции, например ListNet, которая минимизирует кросс-энтропию распределений релевантности и оптимизирует вероятность корректного попадания в топ-k (в контексте решения задачи это топ-1 и топ-3).

Для оценки качества обученных моделей в качестве метрики часто используется *NDCG@3*, отражающая «полезность» первых трёх элементов выдачи с учётом их позиций и релевантности; большее значение метрики качества *NDCG@3* соответствует более точному ранжированию списков. В сравнении с другими видами ранжирования списочный подход обычно обеспечивает более высокую точность, однако требует значительных вычислительных ресурсов. В задачах с небольшим размером списка (шесть элементов) и ограниченным объёмом данных это ограничение, как правило, не является критическим.

Списочный подход демонстрирует высокую устойчивость к шуму в данных: оптимизация сразу всего списка способствует корректной расстановке элементов даже при неточных метках в обучении. Кроме того, такой подход дает возможность внедрения постоянного дообучения: по мере накопления новых оценок кодов Голланда модель можно дообучать на небольших батчах, сохраняя согласованность списков и не теряя ранее выученные связи между метками.

3.1.4. Ансамблевые модели

Для улучшения качества прогноза можно использовать взвешенное ансамблирование моделей (линейный блендинг) [32], когда итоговый ответ вычисляется как линейная комбинация предсказаний различных моделей с оптимизированными коэффициентами. Другим подходом является обучение метамодели на предсказаниях базовых моделей — сте-

кинг.

Для стекинга были взяты следующие базовые модели:

- линейные регрессии с различными параметрами регуляризации;
- L1- и L2-регуляризованные регрессии, LightGBM, CatBoost, случайный лес.

В обоих случаях метамоделью выступает обычная линейная регрессия. Важно отметить, что в первом случае при отсутствии регуляризации получалась бы линейная комбинация линейных моделей, что также является линейной моделью, и именно по этой причине добавляется нелинейная составляющая. Обучение базовых моделей происходит на обучающей выборке, обучение метамодели (подбор весов) — на валидационной, оценка метрик качества — на тестовой.

Для линейного блендинга (взвешенного ансамблирования) требуется найти, с каким весом будет входить в итоговое предсказание каждое из предсказаний базовых моделей. Таким образом, линейная комбинация весов моделей и их предсказания и будет итоговым предсказанием. Применяются следующие подходы для подбора весов [33]:

- Равные веса всех моделей

  каждой базовой модели присваивается одинаковый вклад, что упрощает ансамблирование и служит базовой стратегией;

- Вектор Шэпли [48]

  для каждого возможного порядка добавления моделей в ансамбль вычисляется разница в качестве при включении каждой модели в уже собранный набор, а затем усреднённые маргинальные приросты формируют распределение общего вклада каждого элемента ансамбля;

- Частичный перебор по сетке

  поиск решений на предварительно заданной сетке возможных значений параметров (весов), при увеличении числа моделей для поиска вклада каждой предполагается использование подвыборки заданной сетки;

- Квадратичная оптимизация (QP)

решение задачи минимизации взвешенной суммы ошибок ансамбля как задачи квадратичного программирования с ограничениями;

- Генетический алгоритм (GA)
  эволюционный поиск с выбором лучших представителей популяций (комбинаций весов), их скрещиванием и мутациями;

- Метод роя частиц (PSO) [49]
  оптимизация весов с помощью популяции частиц, которые перемещаются по пространству решений в соответствии с комбинацией собственного и глобального (популяции) оптимального пути;

- Координатный спуск
  итеративная оптимизация веса каждой модели при фиксации остальных.

Стоит отметить, что лишь метод квадратичного программирования предполагает, что функция потерь является непрерывно дифференцируемой. Ансамблирование также служит естественным регуляризатором: при объединении слабых и сильных моделей, уменьшается риск переобучения отдельных компонентов, поскольку ошибки одних моделей компенсируются другими, что особенно ценно при ограниченном объёме данных.

3.2. Математическое обеспечение модуля восстановления данных психометрических тестов

В исходной задаче предполагается наличие пропусков (неполноты) в данных, поскольку пользователь мог и не успеть пройти пять психометрических тестов перед тем, как запрашивается предсказание его кода Голланда. Игнорирование таких записей приведёт либо к значительному сужению выборки (при удалении неполных строк), либо к искажению итоговых оценок. Существуют следующие подходы для восстановления (импутации) данных отсутствующих тестов [50]:

1) множественная импутация цепочными уравнениями ($MICE$);

2) низкоранговая матричная аппроксимация (*Matrix Soft Impute*);

3) применение масок для пропущенных значений;

4) взвешенное ансамблирование комбинации заполненных тестов.

Метод MICE (от англ. *Multivariate Imputation by Chained Equations*) заключается в поочерёдном построении для каждого признака с пропусками регрессионной модели на основе остальных переменных; заполнение пропусков выполняется итеративно в несколько циклов до сходимости. При этом на каждом шаге для конкретного признака используются актуализированные значения остальных признаков, что позволяет учитывать их взаимные зависимости и уменьшать смещение оценок. После завершения всех итераций получается несколько «полных» наборов данных, что даёт возможность оценить неопределённость импутации и корректно скорректировать дисперсию итоговых показателей. Однако метод MICE для обработки *каждой отдельной записи* требует наличия значений по всем остальным признакам: модель фактически «дообучается» на лету при каждом новом наблюдении. Кроме того, при сильной корреляции между переменными могут возникать нестабильность процесса итеративной импутации и проблемы со сходимостью алгоритма.

Метод *Soft Impute* (мягкое матричное восстановление) строит низкоранговую аппроксимацию матрицы с пропусками через регуляризованное сингулярное разложение (SVD, *Singular Value Decomposition*). На каждой итерации отсутствующие элементы заполняются текущими оценками, затем матрица подвергается сингулярному разложению на собственные значения и собственные векторы, после чего к собственным значениям применяется мягкое пороговое преобразование: все значения, не превышающие параметр λ, обнуляются, а остальные уменьшаются на величину λ. Повторяя эти шаги до сходимости, алгоритм одновременно минимизирует ошибку восстановления и контролирует ранжирование компонентов через ядерную норму. Подход, лежащий в основе *Soft Impute*, обеспечивает масштабируемость за счёт применения приближённых алгоритмов сингулярного разложения (SVD).

Находит применение и использование бинарных индикаторов пропусков (масок) — это дополнительные признаки, принимающие значе-

ние «1» при отсутствии исходного измерения и «0» в противном случае. Данный приём не восстанавливает значения пропущенных факторов, а лишь информирует модель о факте выпадения данных, что позволяет учитывать механизмы образования пропусков и повышать устойчивость прогнозов при информативном пропуске. Включение масок пропусков способствует обнаружению зависимостей между фактом отсутствия и целевой переменной, однако повышает размерность признакового пространства и может усложнить обучение моделей за счёт роста числа параметров.

Взвешенное ансамблирование (блендинг) по комбинации тестов предполагает построение отдельного ансамбля для каждой из 31 возможной комбинации заполненных пользователем тестов. Для каждой комбинации подбираются веса базовых моделей с учётом только тех признаков, которые доступны в данном случае, что позволяет максимально адаптировать прогноз к фактическим данным. Такой подход позволяет учесть специфические взаимосвязи между тестами и улучшить точность в каждой группе пользователей, однако требует обучения и хранения 31 набора весов, что значительно увеличивает вычислительные затраты и объём памяти. Кроме того, при появлении новых сочетаний тестов необходимо обновлять все ансамбли, что усложняет сопровождение и масштабирование решения.

В зависимости от объёма и характера пропусков оптимальный выбор метода восстановления данных может различаться: при умеренном уровне неполноты и выраженных корреляциях между тестами MICE и низкоранговая матричная аппроксимация обеспечивают более точное восстановление, тогда как маски пропусков и ансамблирование по комбинациям позволяют обойтись без генерации синтетических значений и сохраняют прозрачность модели. В вычислительном плане методы матричной аппроксимации и применения масок пропусков предпочтительны: оба подхода масштабируются за счёт простой векторизированной реализации и минимально увеличивают объём параметров.

3.3. Итоговая структура решения задачи по определению кода Голланда

Решение задачи по предсказанию кода Голланда предполагает нахождение лучших вариантов на каждом из следующих этапов экспериментального исследования:

1. Восстановление данных тестов, которые не были заполнены.
2. Уменьшение размерности входных данных.
3. Тип решения задачи.
4. Подход в зависимости от типа решения задачи.
5. Базовые модели для каждого из подходов.
6. Настройка гиперпараметров моделей.
7. Взвешенное ансамблирование базовых моделей.

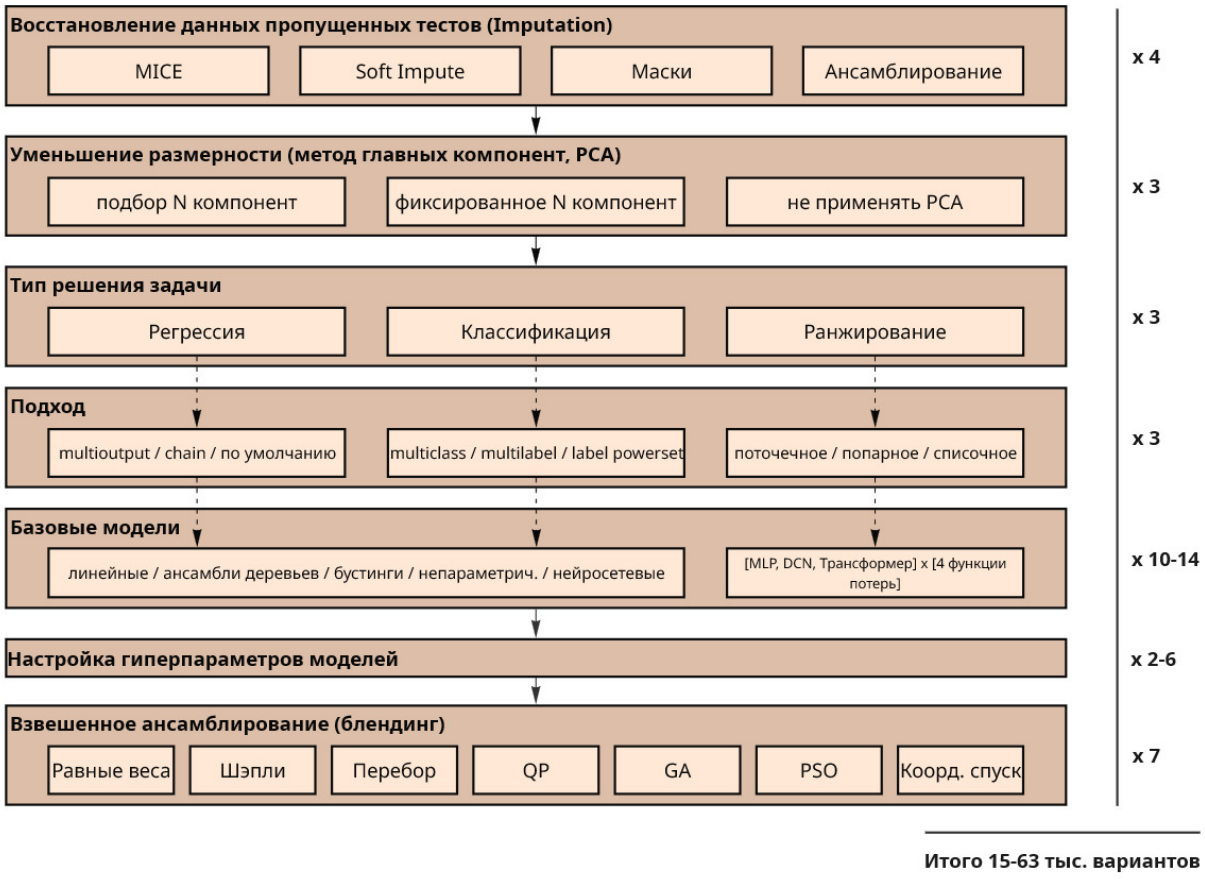

Рисунок 3. Общая схема вариантов экспериментального исследования

Общая схема вариантов экспериментального исследования представ-

лена на рисунке 3. Оценка каждого варианта проводится в соответствии с мерой сходства C-индекс.

Следует отметить, что этап уменьшения размерности может:

- не выполняться вовсе;
- применяться с фиксированным порогом на уровне 90% объяснённой дисперсии;
- выполняться с динамическим подбором числа компонент (N).

На этапе ранжирования поточечный и попарный подходы могут пропускаться в пользу списочного (*listwise*). Конкретный выбор базовых моделей регрессии и классификации более подробно описан в разделах 3.1.1 и 3.1.2. Для списочного ранжирования рассматриваются все возможные комбинации трёх архитектур (многослойный перцептрон, глубокая перекрёстная сеть, трансформер) и четырёх функций потерь (ApproxNDCG, ListNet@1, ListNet@3, LambdaRank). Каждая из базовых моделей также требует подбора гиперпараметров, например, числа ближайших соседей k для модели kNN. Итоговое число вариантов для перебора и определения наилучшего способа предсказания кода Голланда может достигать 63 тысяч различных способов, что создаёт значительную вычислительную нагрузку.

4. Особенности реализации программного инструмента для определения профориентационных предпочтений

В разделе приведены технические детали реализации проекта: используемые языки и библиотеки, описание набора данных, архитектура кода и ключевые функции для подготовки данных, построения и оценки моделей (регрессия, классификация, ранжирование, ансамблирование), а также структура прототипа веб-приложения на R Shiny.

4.1. Общие особенности реализации

Реализация моделей определения профориентационных предпочтений на основе психометрических тестов личности производилась на языке R (версия 4.4.2)[2], являющимся открытым и свободным программным обеспечением. Язык R обеспечивает эффективность обработки данных за счёт векторизованных операций и оптимизированных реализаций статистических алгоритмов в базовых и дополнительных пакетах. Для обработки данных преимущественно использовались библиотеки *data.table* для эффективных табличных преобразований, *tidyverse* (различные манипуляции с данными *dplyr*, элементы функционального программирования с *purrr*, работа со строками через *stringr*), *arrow* для передачи наборов данных между R и Python, визуализация в *plotly*. Нейросетевые модели реализованы на языке Python (версия 3.12.3)[3] с использованием открытого и свободного ПО: *numpy* и *pandas* для численных вычислений и подготовки данных, *PyTorch* как гибкий фреймворк для построения и обучения глубоких нейронных сетей.

Исходный код всего проекта представлен в GitHub-репозитории[4]. В репозитории представлена программная реализация вычислительного эксперимента, представленного на рисунке 3. Далее будут описаны

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. — 2024. URL: https://www.R-project.org/ (дата обращения: 17.05.2025).

[3] Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace. — 2009.

[4] GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/ExP98/Diploma_Holland (дата обращения: 17.05.2025).

ключевые шаги эксперимента с упоминанием названий функций и используемых библиотек, однако описание кода не является исчерпывающим. Код на языке R написан в соответствии с руководством по стилю оформления кода *Tidyverse Style Guide*[5], что обеспечивает единообразие оформления, читаемость и простоту сопровождения.

Инструмент автоматизации процесса профориентации состоит из следующих модулей:

- модуль обработки данных;
- модуль восстановления данных;
- модуль обучения базовых моделей;
- модуль формирования взвешенного ансамбля моделей;
- модуль организации и проведения вычислительных экспериментов.

Предварительно проводился разведочный анализ данных. По результатам вычислительного эксперимента происходил выбор наилучших подходов для определения кода Голланда, обучение и сохранение лучшей модели. На её основе реализуется прототип инструмента автоматизации профориентации: веб-приложение на основе R Shiny.

4.2. Описание набора данных. Разведочный анализ

Данные для исследования были собраны с помощью опросов, размещенных в веб-приложении на базе платформы VK Mini Apps «Психологические тесты»[6] [8]. Приложение находится в открытом доступе и позволяет пользователям проходить различные психометрические опросы, при этом после ознакомления с условиями добровольного информированного согласия[7] пользователи могут разрешить использовать обезли-

[5] Руководство по стилю оформления кода на языке R. Tidyverse Style Guide. URL: https://style.tidyverse.org/ (дата обращения: 17.05.2025).

[6] Мини-приложение «Психологические тесты» (платформа «VK Mini Apps»). URL: https://vk.com/app7794698 (дата обращения: 17.05.2025).

[7] Политика конфиденциальности. URL: https://vk.com/@ticslabs-politika-konfidencialnosti (дата обращения: 17.05.2025).

ченные анонимизированные данные в научных исследованиях (в соответствии с № 152-ФЗ «О персональных данных»).

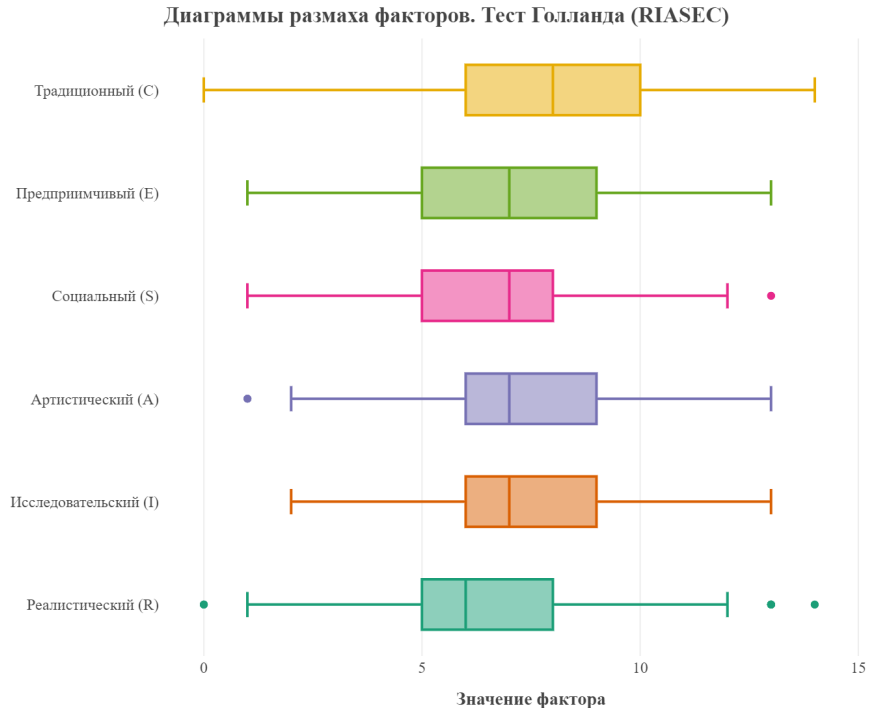

Рисунок 4. Диаграмма размаха факторов модели Голланда

В наборе представлены данные по следующим психометрическим тестам (в скобках указано количество факторов):

1. Опросник Леонгарда-Шмишека (10).
2. Личностный опросник Айзенка (4).
3. 16-факторный опросник Кеттелла (16).
4. Пятифакторный опросник личности («Большая пятёрка»; 5).
5. Ценностный опросник Шварца (20).

Более подробное описание тестов и их факторов представлено в подразделе 2.1.2. Всего имеются данные по 1278 пользователям: у 339 есть данные по всем тестам, у 939 пользователей отсутствует один или два теста. Пример части преобразованного набора данных представлен в таблице 1. Диаграмма размаха для факторов модели Голланда представлена на рисунке 4, плотность их фактических значений — на рисунке 5, частоты значений кодов в порядке убывания их значимости — рисунок 6.

Для анализа линейных связей между факторами модели Голланда на рисунке 7 приведена матрица корреляций: коэффициент корреляции

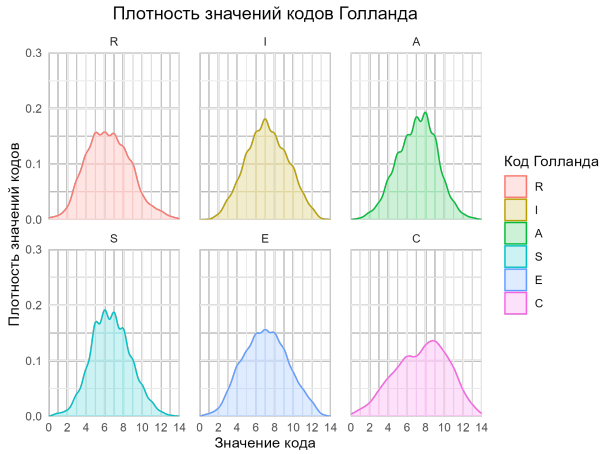

Рисунок 5. Плотность фактических значений кодов Голланда

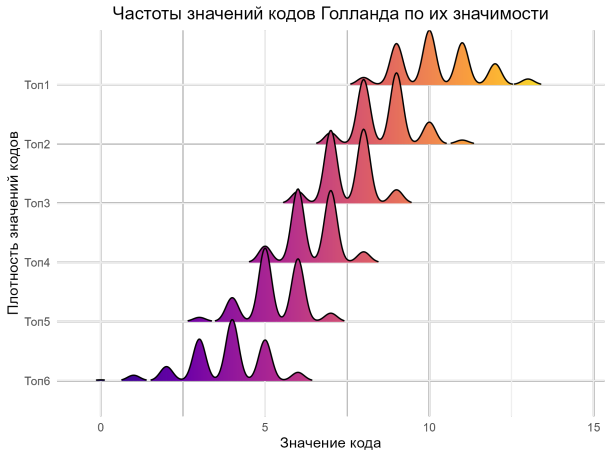

Рисунок 6. Частоты значений кодов Голланда в порядке убывания их значимости

Пирсона по модулю не превышает 0.4 (умереннная зависимость). Описательная статистика по всем факторам всех психометрических тестов приведена в Приложении А в таблице 15. Каждый из представленных факторов является числовой метрической переменной, принимающей только целочисленные значения.

К основным ограничениям исследования относятся особенности сбора данных: возможны смещения из-за специфики портала, а также способа формирования выборки. Для устранения ограничений может быть увеличен размер выборки, включены вопросы о социально-демографических признаках опрашиваемых.

4.3. Особенности реализации модулей программного инструмента

4.3.1. Модуль обработки и восстановления данных

Данные по пройденным пользователями тестам были получены в json-формате, где пользователю поставлен в соответствие пройденный им тест с результатами. Данные были преобразованы в «широкий» табличный формат, заполнены самоочевидные пропуски (исходя из природы психометрических данных), проведена валидация данных в соответствии с допустимыми значениями каждого из факторов психо-

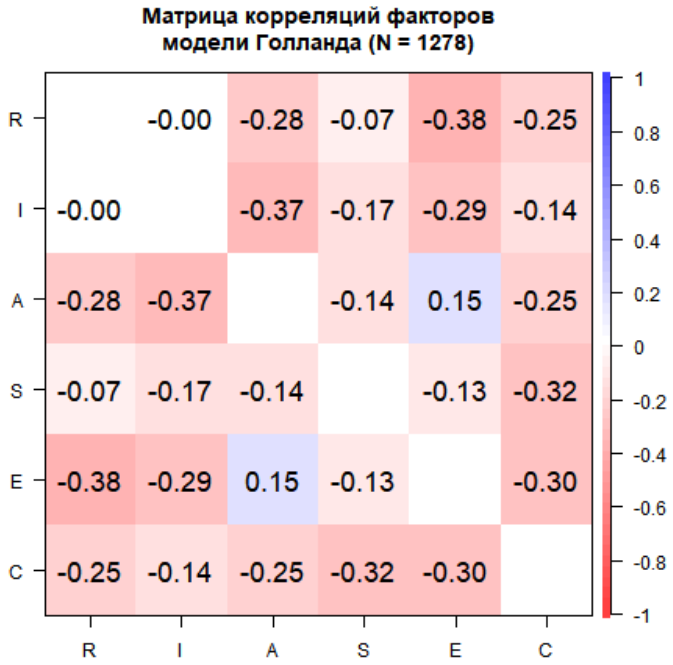

Рисунок 7. Матрица корреляций кодов Голланда

метрических тестов. Для проведения вычислительного эксперимента данные были поделены на обучающую, валидационную и тестовую выборки. Нормализация (стандартизация) данных всех выборок происходит в соответствии со средним значением и с значением стандартного отклонения, вычисляемым по обучающей выборке (функции `transform_data_to_wide`, `train_test_split` в программном коде).

Уменьшение размерности было реализовано в виде метода главных компонент (PCA) с помощью библиотеки *stats*, функции `prcomp`. В результате набора экспериментов было найдено, что чаще всего наилучший результат обеспечивает снижение размерности до 90% объясняемой дисперсии, а именно до 30 компонент; таким образом, сравнение моделей велось и на полном наборе данных, и на данных меньшей размерности ($N_{comp} = 30$). В дальнейшем для моделей (если не указано иное) итоговое значение метрики — это лучшее из значений метрики модели на полном наборе данных и на наборе меньшей размерности.

Восстановление значений факторов незаполненных психометрических тестов для множественной импутации реализовано в функции `mice_imputation` (в основе лежит реализация

Листинг 1. Низкоранговая матричная аппроксимация *Soft Impute*

```
train_matrix_completion <- function(DT, rank.max = 20) {
  aux_feats <- setdiff(colnames(DT), c("id", paste0("HL_", 1:6)))
  M         <- as.matrix(DT)[, aux_feats]
  fit       <- softImpute::softImpute(M, rank.max = rank.max, lambda = 0, type = "als")
  return(list(fit = fit, aux = aux_feats))
}

transform_matrix_completion <- function(fit_obj, DT_new) {
  DT_new <- DT_new %>% as.matrix() %>%
    softImpute::complete(fit_obj$fit) %>%
    {DT_new[, fit_obj$aux] <- .; DT_new}
  return(DT_new)
}
```

из библиотеки *mice*); низкоранговая матричная аппроксимация (листинг 1) — `train_matrix_completion` для учета зависимостей и `transform_matrix_completion` для преобразования с использованием библиотеки *softImpute*; `prepare_mask_data` для применения масок.

4.3.2. Модуль обучения базовых моделей

Базовые модели регрессии и классификации, перечисленные в подразделах 3.1.1 и 3.1.2, в языке R реализованы в различных пакетах: линейная регрессия (*lm*) в пакете *stats*, регуляризованная регрессия — *glmnet*, пошаговая — *MASS*; одноимённые модели в пакетах *xgboost*, *lightgbm*, *catboost*, *randomForest*; метод k-ближайших соседей в *FNN* и *caret*; метод опорных векторов и наивный байесовский классификатор в *e1071*; алгоритм ExtraTrees в *ranger*. Для проведения экспериментов разработан унифицированный интерфейс взаимодействия с моделями: с помощью пакета *R6* реализованы соответствующие классы-обёртки, наследующие шаблонный класс `my_template_model`, в которых при необходимости переопределены методы `initialize()`, `fit()` и `predict()`. Многие модели, например, случайный лес, бустинги, линейные регрессии, позволяют вычислять важность признаков, поэтому для них определен метод `calc_importance()`.

Для решения задачи регрессии с помощью нейронных сетей был взят язык *Python* и фреймворк *PyTorch*. Первая модель из трёх реализованных представляет собой простой многослойный перцептрон с

четырьмя плотными слоями ($55 \to 128 \to 64 \to 32 \to 6$) и функцией активации ReLU: такое снижение размерности позволяет постепенно выделять всё более абстрактные признаки из исходных 55 факторов и сразу выдавать шесть прогнозируемых значений. Вторая модель расширена модулями нормализации (BatchNorm) и регуляризации (Dropout, L1/L2-регуляризация, весовой спад) на каждом скрытом слое, что улучшает сходимость, устойчивость к переобучению и ускоряет обучение: сочетание таких приёмов и глубины должно обеспечивать баланс между выразительностью сети и её обобщающей способностью. В качестве третьей модели используется *TabPFN* (библиотека *tabpfn*) — это трансформер, предобученный на большом числе синтетических табличных задач и предлагающий встроенную байесовскую оценку неопределённости. Он позволяет мгновенно делать предсказания без итеративного обучения, автоматически адаптируясь к различным типам признаков и небольшим объёмам данных. Модель *TabPFN* также демонстрирует стабильные результаты в условиях разнородных и зашумлённых данных.

Решение задачи ранжирования также было реализовано в *Python* с помощью *PyTorch* (подклассы `nn.Module`). Для списочного ранжирования были реализованы три класса:

- `MLPRanker` включает четыре полносвязных слоя размерности $[55 \to 128 \to 64 \to 32 \to 6]$ с последовательным применением BatchNorm и Dropout. Такая глубина позволяет постепенно снижать размерность признаков и выделять всё более абстрактные комбинации без избыточного роста числа параметров.

- `DCNUserItemRanker` сочетает три перекрёстных (*cross*) слоя, моделирующих полиномиальные взаимодействия входных признаков, и «глубокую» ветвь из линейных блоков $[64 \to 128 \to 64 \to 32]$ с ReLU и Dropout. Число перекрёстных слоёв выбрано эмпирически: трёх итераций перекрёстных преобразований достаточно для учёта основных взаимодействий, при этом обучение остаётся устойчивым.

- Модель `ListwiseTransformer` для каждого из шести элементов

создаёт обучаемое представление (эмбеддинг) размерности 8, объединяет его с 55-мерным входным вектором, затем проецирует результат в пространство размерности 64 (`d_model` = 64). Далее он проходит через два слоя кодировщика на основе трансформера с четырьмя головами внимания (`nhead` = 4) и завершается одним линейным выходным слоем. Использование двух слоёв обеспечивает баланс между учётом взаимного влияния элементов и приемлемым временем обучения.

Каждый из классов комбинируется с четырьмя функциями ошибок, соответствующих списочной оптимизации:

- `ListNet@1` и `ListNet@3` вычисляют кросс-энтропию между распределениями рангов (полным или усечённым до топ-3) с помощью функции `softmax`;
- `ApproxNDCG` аппроксимирует классическую метрику NDCG, вводя дифференцируемые ранги через сигмоидные функции;
- `LambdaRank` формирует парно-ориентированную функцию ошибок с расчётом λ-коэффициентов на основе ΔNDCG при перестановках.

4.3.3. Модуль формирования взвешенного ансамбля моделей

Подбор весов моделей для дальнейшего взвешенного ансамблирования (линейного блендинга) был реализован в следующих функциях:

- равные веса всех моделей — функция `equal_weights`;
- вектор Шэпли — `approx_shapley` (листинг 2). Для ускорения работы функции производилась стохастическая аппроксимация с использованием метода Монте-Карло: вместо полного перебора всех вариантов были выбраны R перестановок (разброс значений оценок уменьшается пропорционально $1/\sqrt{R}$);
- частичный перебор по сетке — `grid_search_weights` с заданием шага сетки;

Листинг 2. Функция подсчета важности моделей на основе вектора Шэпли (с применением стохастического метода Монте-Карло)

```
approx_shapley <- function(models_probs, Y_true, R = 500) {
  M <- length(models_probs)
  prob_array <- simplify2array(models_probs)
  phi <- numeric(M)
  for (r in seq_len(R)) {
    perm <- sample.int(M)
    cum_sum <- 0
    v_prev <- 0
    for (k in seq_along(perm)) {
      idx <- perm[k]
      cum_sum <- cum_sum + prob_array[,, idx]
      v_curr <- matr_cind(cum_sum / k, Y_true)
      phi[idx] <- phi[idx] + (v_curr - v_prev)
      v_prev <- v_curr
    }
  }
  w <- phi / R
  return(w / sum(w))
}
```

- квадратичная оптимизация в функции `stacking_qp_weights`, использующая функцию `solve.QP` библиотеки *quadprog*;
- генетический алгоритм и метод роя частиц — реализованы на основе функций `ga` библиотеки *GA* и `psoptim` из *PSO* (код функции `particle_swarm_weights` приведен в листинге 3);
- координатный спуск — функция `coordinate_optimize_weights` (использует библиотеку *stats*).

От использования подбора весов с помощью байесовской оптимизации (библиотека *rBayesianOptimization*) было решено отказаться вследствие больших вычислительных затрат (разница во времени выполнения, например, с методом роя частиц в два порядка).

Таким образом, способы подбора весов были реализованы как собственными средствами, так и с использованием возможностей, встроенных в различные пакеты.

Листинг 3. Функция подсчета важности моделей на основе метода роя частиц (PSO)

```
particle_swarm_weights <- function(models_probs, Y_true, swarm_size = 50, maxit = 100) {
  M <- length(models_probs)
  fn_pso <- function(x) {-weighted_cindex_value(x / sum(x), models_probs, Y_true)}
  PSOres <- pso::psoptim(par = rep(1/M, M), fn = fn_pso, lower = rep(.Machine$double.eps,
  ↪  M), upper = rep(1, M), control = list(s = swarm_size, maxit = maxit, trace = FALSE,
  ↪  vectorize = TRUE, maxit.stagnate = 10))
  w <- PSOres$par
  w[w < 1e-6] <- 0
  return(w / sum(w))
}
```

4.3.4. Модуль организации и проведения вычислительных экспериментов

При разработке решения задачи многоцелевой регрессии были реализованы походы независимого предсказания выходов (значений кодов Голланда) `stack_MO_regr`, предсказаний по цепочке `chain_MO_regr` и `no_MO_regr` для предсказаний моделей, поддерживающих множественные выходы по умолчанию. Наличие унифицированных интерфейсов (R6-классов) с базовыми моделями позволило организовать массовую оценку моделей с помощью функции `calc_regression_models`, на вход которой подается следующий набор данных: название класса модели, её метка и гиперпараметры, подход к многоцелевой регрессии; код представлен в листинге 4. Аналогичный подход был использован для реализации подходов к решению задачи классификации: на основе унифицированной функции-интерфейса `classification_test_framework`, позволяющего обучить модель, сделать предсказание, оценить значения метрик качества (C-индекс и точность Top-K), применяется функция, например, `run_multilabel_experiments` для многометочной классификации (листинг 5).

Листинг 4. Функция массовой оценки регрессионных моделей

```
calc_regression_models <- function(regr_models_df, X_train_, Y_train_, X_test_, Y_test_) {
  MO_res <- regr_models_df %>%
    copy() %>%
    .[, pred := pmap(list(model, params, MO_type), \(mdl, par, mo_type)
        do.call(perform_MO_regression,
          c(list(mdl, mo_type, X_train_, Y_train_, X_test_, Y_test_, print_model_name = F),
            par)))] %>%
    .[, rmse := map_dbl(pred, \(x) df_metric(x, Y_test_, my_rmse) %>% round(3))] %>%
    .[map_lgl(pred, \(item) !is.null(item)),
      C_index := map_dbl(pred, \(x) df_metric(x, Y_test_, calc_C_index) %>% round(3))]
      %>%
    .[, .(name, pred, rmse, C_index)]
  return(MO_res)
}
```

Листинг 5. Функция массовой оценки решения задачи многометочной классификации

```
run_multilabel_experiments <- function(experiments_df, X_train, Y_b_train,
                                       X_test, Y_b_test) {
  # experiments_df (tibble/data.table): clsf_func, label, params, n_retry
  evaluate_ML <- function(multlbl_clsf_func, label = "", n_retry = 1, ...) {
    classification_test_framework(Y_test = Y_b_test, Y_b_test = Y_b_test,
                                  clsf_func = multlbl_clsf_func, n_retry = n_retry,
                                  label = label, X_train = X_train,
                                  Y_b_train = Y_b_train, X_test = X_test, ...)
  }

  res <- experiments_df %>%
    as.data.table() %>%
    .[, metrics := pmap(list(multlbl_clsf_func, label, params, n_retry), \(ind_fn, lbl,
      extra_args, nr)
                        do.call(evaluate_ML, c(list(multlbl_clsf_func = ind_fn, label =
                          lbl, n_retry = nr), extra_args))
    )] %>%
    .[, .(label, metrics, id = 1:.N)] %>%
    .[, metrics[[1]], by = id]

  return(res)
}
```

4.4. Реализация прототипа инструмента для определения профориентационных предпочтений

Прототип инструмента для определения профориентационных предпочтений был реализован в виде интерактивного веб-приложения на платформе R Shiny. На рисунке 8 приведён общий порядок вычислительного конвейера разработанного прототипа программного инструмента, в котором последовательно выполняются все необходимые шаги: от предобработки и очистки исходных данных до восстановления пропусков с помощью модели мягкого матричного восстановления *Soft Impute* и дальнейшего прогнозирования с помощью предварительно обученной и сохраненной модели регуляризованной L1-регрессии (*Lasso*), показавшей наилучшие результаты среди базовых моделей (при оценке C-индекса).

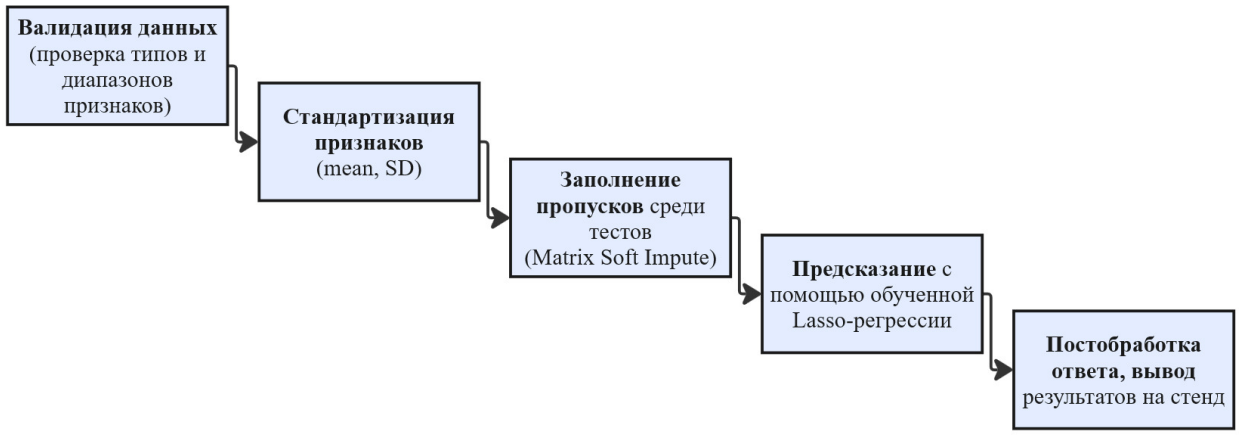

Рисунок 8. Итоговая последовательность вычислительных шагов

При разработке веб-стенда на R Shiny в блоке кода `ui` через `fluidPage` и `bsCollapse` с помощью функции `create_test_ui` динамически формируется панель тестов с групповыми сворачивающимися блоками вопросов. В серверной части с помощью `reactiveValues` для хранения промежуточных данных и `observeEvent(input$calc)` организована реактивная логика: при нажатии «Подсчитать» происходит валидация входных значений (в случае выхода за рамки допустимых значений пользователю выводится сообщение об ошибке), восстановление пропусков методом *Soft Impute* и нормализация данных, затем вызывается функция предсказания, а результаты выводятся через

`renderUI` и подтверждаются уведомлениями `showNotification`. По завершении вычислений результаты в виде набора кодов рекомендуемых направлений сопровождаются подробным текстовым описанием и визуальной презентацией на пользовательском интерфейсе (см. интерфейс стенда на рисунках 9 и 11). Разделение на `ui` и `server` типично для *Shiny*-приложений: `ui` описывает компоненты интерфейса, а сервер — реактивную логику и обновление данных. Такой подход обеспечивает раздельную ответственность за внешний вид и вычислительную логику прототипа.

Доступ к приложению не требует процедуры авторизации, при этом его развертывание возможно как в локальной среде, так и на удалённом сервере с использованием средств хостинга. Разработанный прототип может быть использован в составе более комплексных информационных систем, направленных на поддержку профессиональной ориентации и образовательного планирования.

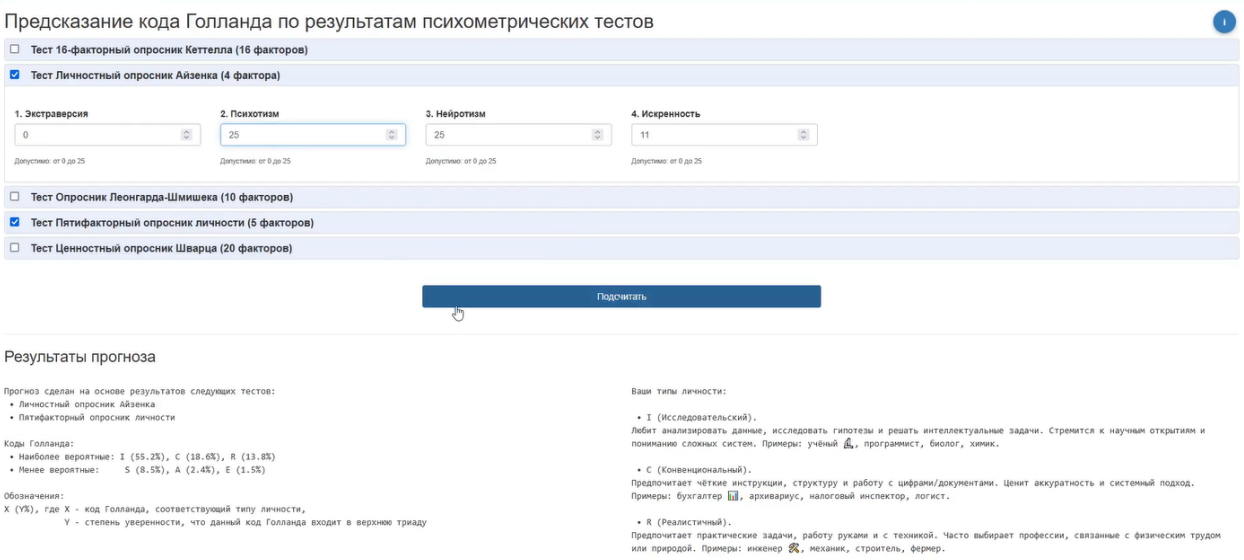

Рисунок 9. Интерфейс прототипа инструмента профориентации

5. Результаты вычислительного эксперимента по определению кода Голланда

5.1. Сравнение результатов моделей на полных данных

Для обеспечения сравнимости различных подходов (регрессия, классификация, ранжирование) между собой основное внимание было уделено мере сходства C-индекс (более подробное описание в подразделе 2.1.1; чем выше значение, тем лучше; значение для случайного константного предсказателя равно 9.0). Такой выбор меры позволяет сравнивать результаты с другими научными работами. Так, на основе социо-демографических данных в работе [14] достигается значение C-index = 10.95 для регрессии и C-index = 11.08 для классификации, поэтому результат предсказания с таким же значением или больше может считаться приемлемым.

Результаты определения кода Голланда как задачи регрессии (на основе C-индекса) приведены в таблице 3. В сравнении с константным предсказателем модель XGBoost показывает низкие результаты. Наилучшие результаты показывает регуляризованная регрессия (Lasso и Ridge): C-index = 11.175 и C-index = 11.062. Высокий результат показывает модель ExtraTrees: C-index = 11.1. Лучшая из нейросетевых моделей, foundation-модель TabPFN, C-index = 10.562, показывает результат лучше лишь XGBoost и на одном уровне с методом k-ближайших соседей. Стоит отметить, что результаты моделей при различных подходах, независимом и по цепочке (*multioutput* и *chain*), практически идентичны. В то же время попарно для каждого из подходов лучшие результаты многие модели показывают на наборе данных меньшей размерности (после применения метода главных компонент, PCA), кроме регуляризованных линейных регрессий.

На примере модели случайного леса (*Random Forest*) в таблице 4 приведен анализ важности признаков. Так, два фактора из модели Кеттелла покрывают более 30% важности всех 55 факторов, 8 факторов — более 50%. В Random Forest важность признака (*gain*) — это усреднён-

Таблица 3. Сравнение регрессионных моделей, C-индекс

Модель	Multi-output	Mult. PCA	Chain	Chain PCA
Регрессия Lasso (L1)	11.175	10.887	11.175	11.150
ExtraTrees	10.700	11.100	10.625	10.825
Регрессия Ridge (L2)	10.988	10.537	11.062	10.412
Метод опорных векторов	10.713	10.950	10.713	10.950
Пошаговая регрессия	10.605	10.905	10.600	10.905
CatBoost	10.688	10.812	10.688	10.812
Случайный лес	10.625	10.475	10.812	10.588
Линейная регрессия (OLS)	10.688	10.800	10.688	10.800
LightGBM	10.750	10.425	10.750	10.425
kNN	10.525	10.400	10.525	10.400
XGBoost	9.164	9.729	9.162	9.725
Constant baseline	9.000	9.000	9.000	9.000
TabPFN	10.562			
MLP (BatchNorm, DropOut, регуляризация)	10.462			
MLP	10.275			

Обозначения:
Mult. — multioutput, предсказание переменных независимо друг от друга,
Chain — предсказание выходных переменных по цепочке,
PCA — метод главных компонент (уменьшение размерности)

ная по всем деревьям сумма уменьшений критерия нечистоты (Джини или энтропии) на узлах, где при разбиении использовался этот признак. При аналогичном анализе важности признаков с помощью моделей градиентного бустинга получаются схожие результаты: наиболее важными признаются схожие признаки, но они имеют меньшую важность.

Сравнение методов подбора весов ансамбля регрессионных моделей представлено в таблице 5. Лучшим методом подбора весов для моделей линейного блендинга (взвешенного ансамблирования) является метод роя частиц (PSO), C-index $=$ 11.663. В таблице 6 приведены веса входящих в лучшую PSO-модель базовых регрессоров: наибольший вклад вносит Lasso-регрессор (43.2%), а также пошаговая регрессия. Модели стекинга показывают результаты хуже, чем модели линейного блендинга.

Таблица 4. Важность признаков модели случайного леса

Код признака	Наименование признака	Важность (%)	Накоплено (%)
CT_1	Открытость–замкнутость	15.5	15.5
CT_7	Чувственность–твердость	15.5	31.0
SC_19	Гедонизм–индивидуальный приоритет	4.2	35.2
EY_1	Экстраверсия	4.0	39.2
CT_4	Беспечность–озабоченность	3.6	42.8
SC_3	Власть–нормативный идеал	3.4	46.2
LN_3	Циклотимность	3.3	49.5
BF_3	Самоконтроль–импульсивность	2.5	52.0
BF_4	Эмоц. устойчивость–неустойчивость	2.5	54.5

Таблица 5. Сравнение методов подбора весов ансамбля регрессионных моделей

Метод подбора весов	**Multi-output**	**Mult. топ-5**	**Chain**	**Chain топ-5**
Равные веса всех моделей	11.063	11.088	11.050	11.013
Вектор Шэпли (Shap)	11.050	11.138	11.138	11.050
Частичный перебор по сетке	11.550	11.388	11.538	11.325
Квадратичная оптимизация (QP)	10.588	10.463	10.738	10.813
Генетический алгоритм (GA)	11.500	11.550	11.300	11.563
Метод роя частиц (PSO)	11.600	11.663	11.613	11.613
Координатный спуск	11.188	11.225	11.288	11.413
Линейные регрессии с регуляризацией	Линейная регр.		10.887	
Lasso, Ridge, LightGBM, CatBoost, RF	Линейная регр.		10.688	

Обозначения:
Mult. — Multioutput, топ-5 — подбор весов только для топ-5 моделей по C-индексу

Таблица 6. Весовые коэффициенты моделей и C-индекс для PSO

Подбор	**Веса моделей**				**C-**
весов	Lasso L1	Пошаговая регр.	CatBoost	ExtraTrees	**индекс**
PSO	0.432	0.327	0.150	0.091	11.663

Таблица 7. Сравнение подходов к классификации (Top-K accuracy)

Модель	**Multiclass**			**Multilabel**			**Label Powerset**		
	Top1	Top2	Top3	Top1	Top2	Top3	Top1	Top2	Top3
kNN	0.99	0.71	0.13	1.00	0.76	0.11	0.98	0.65	0.18
Логист. L1-регр.	1.00	0.70	0.16	1.00	0.70	0.16	0.99	0.64	0.10
XGBoost	1.00	0.70	0.11	0.98	0.68	0.10	0.96	0.63	0.11
Логист. L2-регр.	1.00	0.70	0.15	0.99	0.70	0.21	0.99	0.68	0.09
Наивный Байес	0.98	0.70	0.15	0.99	0.70	0.15	0.99	0.69	0.16
ExtraTrees	1.00	0.73	0.15	1.00	0.78	0.15	0.98	0.69	0.20
SVM	1.00	0.74	0.15	1.00	0.72	0.14	0.98	0.68	0.21
Random Forest	1.00	0.74	0.16	1.00	0.74	0.15	0.99	0.64	0.23
CatBoost	0.99	0.79	0.11	0.99	0.79	0.11	0.99	0.70	0.16
LightGBM	0.98	0.56	0.05	0.98	0.70	0.10	0.95	0.53	0.05

На рисунке 10 показана гистограмма распределения значений C-индекса для предсказаний PSO-ансамбля регрессоров. Заметно, что распределение «скошено» влево относительно медианы; большая часть значений лежит правее константного значения C-index $= 9.0$.

Сравнение моделей базовых классификаторов в разрезе трех главных подходов для классификации показано в таблице 7. Сравнение производилось по метрике Top-K точность. Подходы *multiclass* и *multilabel* показывают схожие результаты и опережают подход *label powerset* по метрикам точности Top-1 и Top-2. Можно заметить, что большинство моделей в 98–100% случаев верно предсказывают в тройке кодов хотя бы один код, который действительно есть в фактических данных в *верхней триаде*; в 70% и более угадываются хотя бы два кода. Лишь примерно в 15% правильно предсказываются все три кода одновременно.

В таблице 8 приведено сравнение лучших базовых моделей класси-

Таблица 8. Сравнение классификаторов

Классификатор	Подход	C-индекс	Top1	Top2	Top3
kNN	Multilabel	10.838	1.000	0.763	0.113
Логист. L1-регр.	Multiclass	10.663	1.000	0.700	0.163
XGBoost	Multiclass	10.638	1.000	0.700	0.113
Логист. L2-регр.	Multiclass	10.500	1.000	0.700	0.150
Наивный Байес	Multilabel	10.350	0.988	0.700	0.150
ExtraTrees	Multilabel	10.013	1.000	0.775	0.146
SVM	Multilabel	9.875	1.000	0.721	0.138
Random Forest	Multilabel	9.800	0.996	0.738	0.146
CatBoost	Multilabel	9.775	0.988	0.788	0.113
LightGBM	Multilabel	9.313	0.975	0.700	0.100
Baseline (случ.)	–	9.000	0.950	0.500	0.050

Таблица 9. Сравнение методов подбора весов ансамбля классификаторов

Метод подбора весов	Multiclass	Multilabel	Label Powerset
Равные веса всех моделей	10.663	10.888	10.563
Вектор Шэпли (Shap)	10.563	11.038	10.525
Частичный перебор по сетке	11.213	11.488	11.525
Квадратичная оптимизация (QP)	10.488	10.638	10.650
Генетический алгоритм (GA)	11.263	11.313	11.213
Метод роя частиц (PSO)	11.263	11.625	11.525
Координатный спуск	11.200	11.275	10.425

фикации. На первом месте с C-index = 10.838 метод k-ближайших соседей (подход *multilabel*), за ним логистическая Lasso-регрессия (*multiclass*) с C-index = 10.663.

Сравнение методов подбора весов ансамбля классификаторов представлено в таблице 9. Лучшим методом подбора весов для моделей линейного блендинга (взвешенного ансамблирования) является метод роя частиц (PSO), C-index = 11.625. В таблице 10 приведены веса классификаторов, которые определены с помощью PSO: наибольший вклад вносят kNN и L1-регрессия. Таким образом, результаты лучшего взвешенного ансамбля классификаторов лишь немного хуже лучшего

Рисунок 10. Распределение значений C-индекса для предсказаний PSO-ансамбля регрессоров

Таблица 10. Весовые коэффициенты моделей и C-индекс для PSO

Подбор	**Веса моделей**					**C-**
весов	kNN	SVM	Logit L1	XGBoost	LightGBM	**индекс**
PSO	0.291	0.164	0.191	0.183	0.151	11.625

Таблица 11. Сравнение моделей ранжирования

Функция потерь	C-индекс			NDCG@3		
	Deep & Cross	Trans-former	MLP	Deep & Cross	Trans-former	MLP
ApproxNDCG	10.025	8.888	9.150	0.539	0.439	0.388
LambdaRank	9.963	9.675	9.650	0.527	0.489	0.543
ListNet@1	9.650	10.325	10.438	0.504	0.628	0.653
ListNet@3	9.450	9.950	10.788	0.458	0.622	0.638

регрессионного ансамбля.

Сравнение моделей ранжирования отражено в таблице 11. Лучшая из моделей, многослойный перцептрон с функцией потерь ListNet@3, показывает C-index = 10.788, что заведомо хуже, чем лучшие из базовых регрессоров и классификаторов.

Таблица 12. Обзор лучших моделей для каждого типа задач

Тип задач	Подход	Лучшая модель	C-индекс
Регрессия	Блендинг, mo	PSO (L1-регрессия, пошаговая регрессия, CatBoost, ExtraTrees)	11.663
Классификация	Блендинг, ml	PSO (kNN, SVM, логистич. L1-регр., XGBoost, LightGBM и др.)	11.625
Регрессия	Блендинг, chain	PSO	11.613
Классификация	Блендинг, lp	PSO / поиск по сетке	11.525
Классификация	Блендинг, mc	Генетический алг. / PSO	11.263
Регрессия	Multioutput	L1-регрессия	11.175
Регрессия	Chain	L2-регрессия	11.062
Классификация	Multilabel	kNN	10.838
Ранжирование	Списочное	MLP с ListNet@3	10.788
Классификация	Multiclass	Логистич. L1-регрессия	10.663

Обозначения:
mo — Multioutput, ml — Multilabel, lp — Label Powerset, mc — Multiclass,
MLP — многослойный перцептрон, PSO — метод роя частиц,
SVM — метод опорных векторов, kNN — метод k-ближайших соседей

Лучшие из моделей по каждому рассматриваемому типу задач приведены в таблице 12. Двумя лучшими моделями оказались линейные блендинги, веса которых подобраны методом роя частиц: это подходы *multioutput* для регрессии (C-index = 11.663) и *multilabel* для классификации (C-index = 11.625). Результаты базовых моделей уступают результатам их комбинаций в виде взвешенных ансамблей.

5.2. Сравнение подходов к восстановлению данных тестов

Результаты предсказания регрессионных моделей с использованием различных подходов к восстановлению значений незаполненных психометрических тестов представлены в таблице 13. Лучшие базовые модели для восстановления результатов — это Lasso- и пошаговый регрессоры в сочетании с подходом мягкого матричного восстановления данных (C-index = 10.248 и C-index = 10.183). Ансамблирование моделей, обученных на данных заполненных тестов по отдельности, затем объединенных воедино в зависимости от наличия того или иного теста, по-

Таблица 13. Восстановление значений незаполненных психометрических тестов

Модель-регрессор	**MICE**	**Soft Impute**	**Маски**	**Ансамбли**
Регрессия Lasso (L1)	9.191	10.248	9.998	9.866
Пошаговая регрессия	9.608	10.183	9.978	10.082
Random Forest	9.518	10.136	9.819	9.712
LightGBM	9.372	10.086	9.686	9.594
Линейная регр. (OLS)	9.407	10.021	9.876	10.012
Регрессия Ridge (L2)	9.442	9.770	9.868	9.933
ExtraTrees	9.101	9.823	9.870	9.808
Метод опорных векторов	9.221	9.814	9.864	9.760
CatBoost	9.131	9.814	9.835	9.461
kNN	9.372	9.834	9.830	9.377
XGBoost	8.769	9.571	9.267	9.614
Constant baseline	9.000	9.000	9.000	9.000

Таблица 14. Весовые коэффициенты моделей и C-индекс при разных методах подбора весов для ансамблей на восстановленных данных

Подбор	**Веса моделей**					**C-**
весов	**Lasso L1**	**Пошаг.**	**LightGBM**	**Случ. лес**	**kNN**	**индекс**
PSO	0.001	0.481	0.038	0.475	0.005	10.740
Grid	0.000	0.500	0.000	0.500	0.000	10.657
GA	0.281	0.369	0.109	0.189	0.052	10.401
Спуск	0.019	0.422	0.067	0.305	0.187	10.245
QP	0.050	0.000	0.390	0.007	0.553	10.065
Равные	0.200	0.200	0.200	0.200	0.200	10.053
Шэпли	0.247	0.185	0.206	0.179	0.183	10.047

Обозначения:
PSO — метод роя частиц, Grid — частичный перебор по сетке,
GA — генетический алгоритм, спуск — координатный спуск,
QP — квадратичная оптимизация, пошаг. — пошаговая регрессия,
случ. лес — случайный лес (Random Forest), kNN — метод k-ближайших соседей

казывает следующие результаты: наибольшее значение C-индекса для пошаговой моделей регрессии лишь немного хуже подхода *Soft Impute*: C-index = 10.082, однако ансамблевый подход по отдельным тестам требует обучения для всех 31 комбинации наличия тестов. Применение масочного подхода к восстановлению данных в меньшей степени зависит от выбора базовой модели. Хуже всего себя показывает подход множественной импутации (*MICE*).

В таблице 14 был применен лучший из подходов к восстановлению данных, мягкое матричное восстановление, и проведен вычислительный эксперимент с регрессионными моделями, как если бы данные были полными. Для ансамблирования были взяты топ-5 моделей, показавших наибольшие значения меры сходства на этих данных по отдельности. Наилучший результат достигнут при подборе весов для ансамбля с помощью метода роя частиц (*PSO*): C-index = 10.74, где наибольший вклад вносят модели пошаговой регрессии (48.1%) и случайного леса (47.5%). Таким образом, ансамблирование позволяет значительно улучшить результаты предсказания на восстановленных данных, однако значения меры сходства на восстановленных данных ниже, чем на полных данных для аналогичных моделей регрессии и классификации.

Заключение

В ходе работы была достигнута поставленная цель: разработан инструмент для автоматизации профориентации на основе предсказания кода Голланда по неполным результатам психометрических тестов личности. Для выполнения цели были решены следующие задачи:

1. Разработаны и реализованы математические модели модуля восстановления пропусков результатов психометрических тестов: *MICE*, маски, ансамбли по набору заполненных тестов и метод мягкой импутации, который в сочетании с PSO-ансамблем регрессоров показал наибольший C-индекс: 10.74.

2. Реализованы подходы к определению кода Голланда: многоцелевая регрессия (*multioutput*, *chain*), классификация (*multiclass*, *multilabel*, *label powerset*), списочное ранжирование.

3. Разработан модуль формирования взвешенного ансамбля моделей на основе методов подбора весов: метод роя частиц (PSO), равные веса, частичный перебор по сетке, вектор Шэпли, квадратичная оптимизация, генетический алгоритм и координатный спуск.

4. Проведен сравнительный анализ моделей предсказания кодов Голланда; лучшие результаты у моделей линейного блендинга с оптимизацией весов моделей методом роя частиц:
   - ансамбль *multioutput*-регрессоров: Lasso- и пошаговая регрессии, CatBoost, ExtraTrees (C-index = 11.663);
   - ансамбль *multilabel*-классификаторов: kNN, SVM, логистическая Lasso-регрессия, XGBoost, LightGBM и др. (C-index = 11.625);
   - лучшая базовая модель — L1-регрессия со множественными выходами (C-index = 11.175);
   - показано превосходство классических методов машинного обучения над нейросетевыми в данной задаче.

5. Создан прототип инструмента для определения профориентационных предпочтений на основе R Shiny.

Исходный код всего проекта представлен в GitHub-репозитории[8]. Отдельные аспекты вычислительного эксперимента, связанные с решением задач регрессии и классификации, ранее были представлены на XXVIII Международной конференции по мягким вычислениям и измерениям SCM'2025.

[8] GitHub: Предсказание кода Голланда (RIASEC) по результатам психометрических тестов личности. URL: https://github.com/ExP98/Diploma_Holland (дата обращения: 17.05.2025).

Список литературы

[1] Lo Presti Alessandro, Capone Vincenza, Aversano Ada, Akkermans Jos. Career Competencies and Career Success: On the Roles of Employability Activities and Academic Satisfaction During the School-to-Work Transition // Journal of Career Development. – 2021. – Vol. 49.

[2] Aydıntan Belgin, Koç Hakan. The Relationship between Job Satisfaction and Life Satisfaction: An Empirical Study on Teachers // International Journal of Business and Social Science. – 2016. – Vol. 7.

[3] Cannas Massimo, Sergi Bruno, Sironi Emiliano, Mentel Urszula. Job satisfaction and subjective well-being in Europe // Economics and Sociology. – 2019. – Vol. 12. – P. 183–196.

[4] Medgyesi Marton, Zolyomi Eszter. Job satisfaction and satisfaction in financial situation and their impact on life satisfaction. Social Situation Monitor Research Note 6/2016 // Directorate-General for Employment, Social Affairs and Inclusion 2016. – 2016.

[5] Федеральный закон от 30 декабря 2020 г. № 489-ФЗ "О молодежной политике в Российской Федерации". – Собрание законодательства Российской Федерации, 2021, № 1, ст. 1. – 2020.

[6] Pordelan Nooshin, Hosseinian Simin. Design and development of the online career counselling: a tool for better career decision-making // Behaviour and Information Technology. – 2020. – Vol. 41. – P. 1–21.

[7] Westman S., Kauttonen J. et al. Artificial Intelligence for Career Guidance – Current Requirements and Prospects for the Future // IAFOR Journal of Education. – 2021. – Vol. 9. – P. 43–62.

[8] Лаборатория прикладного искусственного интеллекта СПб ФИЦ РАН. Мини-приложение «Психологические тесты». – URL: https://vk.com/app7794698 (дата обращения: 22.05.2025).

[9] Holland J. L. A theory of vocational choice // Journal of Counseling Psychology. – 1959. – Vol. 6, no. 1. – P. 35–45.

[10] В. Резапкина Г. Психология и выбор профессии: программа предпрофильной подготовки. Учебно–методическое пособие для психологов и педагогов. – Генезис, М., 2005.

[11] Chu Chu, Russell Mary, Hoff Kevin et al. What Do Interest Inventories Measure? The Convergence and Content Validity of Four RIASEC Inventories // Journal of Career Assessment. – 2022.

[12] Hoff Kevin at al. Interested and employed? A national study of gender differences in basic interests and employment // Journal of Vocational Behavior. – 2024. – Vol. 148.

[13] Nye Christopher. Assessing Interests in the Twenty-First-Century Workforce: Building on a Century of Interest Measurement // Annual Review of Organizational Psychology and Organizational Behavior. – 2022. – Vol. 9.

[14] Bogacheva Eugenia, Tatarenko Filipp, Smetannikov Ivan. Predicting Vocational Personality Type from Socio-demographic Features Using Machine Learning Methods // International Conference on Control, Robotics and Intelligent System. – 2020. – P. 93–98.

[15] Chekalev A., Khlobystova A., Abramov M. Community Theme Analyser: Predicting Career Guidance in Online Social Networks // Proceedings of the Eighth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'24), Volume 2. – Cham : Springer Nature Switzerland, 2024. – P. 153–162.

[16] Ivaschenko Anastasia, Abramov Maxim, Stoliarova Valerie. Exploring the possibility of predicting users' career guidance preferences based on analysis of community topics and the gender in the online social network users' profiles // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. – 2023. – Vol. 23. – P. 564–574.

[17] Stoliarova Valerie, Bushmelev Fedor, Abramov Maxim. Associations between the Avatar Characteristics and Psychometric Test Results of VK Social Media Users // Mathematics. – 2023. – Vol. 11.

[18] Oliseenko Valerii, Ivaschenko Anastasia, Korepanova A., Tulupyeva T. Automating the Temperament Assessment of Online Social Network Users // Doklady Mathematics. – 2024. – Vol. 108.

[19] Stankevich Maksim, Ignatiev Nikolay, Smirnov Ivan, Kiselnikova Natalia. Personality Traits Prediction from Vkontakte Social Media // Voprosy kiberbezopasnosti. – 2019. – P. 80–87.

[20] Titov Sergey, Novikov Pavel, Mararitsa Larisa. Full-scale Personality Prediction on VKontakte Social Network and its Applications // 2019 25th Conference of Open Innovations Association. – 2019. – P. 317–323.

[21] Başaran Seren, Ejimogu Obinna. A Neural Network Approach for Predicting Personality From Facebook Data // Sage Open. – 2021.

[22] Mason Rod, Roodenburg John. Personality and vocational interest typologies associated with better coping and resilience in paramedicine: A review of two models // Paramedicine. – 2023. – Vol. 21.

[23] Rúa Sandra M. Hurtado, Stead Graham B., Poklar Ashley E. Five-Factor Personality Traits and RIASEC Interest Types: A Multivariate Meta-Analysis // Journal of Career Assessment. – 2019. – Vol. 27, no. 3. – P. 527–543.

[24] Batista Jonatan, Guedes-Gondim Sonia. Personality and Person-Work Environment Fit: A Study Based on the RIASEC Model // International Journal of Environmental Research and Public Health. – 2022. – Vol. 20. – P. 719.

[25] Usslepp Nele, Hübner Nicolas, Stoll Gundula et al. RIASEC Interests and the Big Five Personality Traits Matter for Life Success—But Do They Already Matter for Educational Track Choices? // Journal of Personality. – 2020. – Vol. 88.

[26] Schuerger J. Career Assessment and The Sixteen Personality Factor Questionnaire // Journal of Career Assessment. – 1995. – Vol. 3. – P. 157–175.

[27] Yamashita Jumpei, Iwai Ritsuko, Oishi Haruo, Kumada Takatsune. Personality Traits Systematically Explain the Semantic Arrangement of Occupational Preferences // Journal of Individual Differences. – 2024. – Vol. 45. – P. 201–217.

[28] Martins Gustavo at al. Assessment of Vocational Interests by Areas of Psychology: Relations with the Big Five and RIASEC // Trends in Psychology. – 2024.

[29] Silva Amila. JPLink: On Linking Jobs to Vocational Interest Types // Advances in Knowledge Discovery and Data Mining. – Cham : Springer International Publishing, 2020. – P. 220–232.

[30] Song Q. Chelsea, Shin Hyun Joo, Tang Chen et al. Investigating machine learning's capacity to enhance the prediction of career choices // Personnel Psychology. – 2022. – Vol. 77.

[31] Bishop Christopher M. Pattern Recognition and Machine Learning. – New York : Springer, 2006. – P. 140–155. – ISBN: 978-0-387-31073-2. – Chapter on Linear Models for Regression.

[32] Zhang C., Ma Y. Ensemble machine learning: Methods and applications. – 2012. – P. 1–329.

[33] Bischl Bernd, Binder Martin, Lang Michel, Pielok. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2023. – Vol. 13.

[34] Schmieschek H. G. Questionnaire for the determination of accentuated personalities // Psychiatrie, Neurologie, und medizinische Psychologie. – 1970. – Vol. 22 10. – P. 378–81.

[35] Кортнева Ю. В. Диагностика актуальной проблемы: методика Леонгарда–Шмишека. – Москва : Институт общегуманитарных исследований, 2004. – 237 с. – ISBN: 5-88230-151-3.

[36] Eysenck Hans, Eysenck Sybil, Barrett Paul. A revised version of the Psychoticism scale // Personality and Individual Differences. – 1985. – Vol. 6. – P. 21–29.

[37] Суходольский Г. В. Математическая психология. – Харьков : Гуманитарный Центр, 2006. – 358 с. – ISBN: 966-8324-24-2.

[38] Cattell Heather, Mead Alan. The Sixteen Personality Factor Questionnaire (16PF) // Sage Publications. – 2008.

[39] Шмелёв А. Г. Психодиагностика личностных черт. – Санкт-Петербург : Речь, 2002. – 480 с. – ISBN: 5-9268-0084-6.

[40] Tsuji Heijiro, Fujishima Yutaka. Five-factor model of personality: Concept, structure, and measurement of personality traits // Japanese Psychological Review. – 1997. – Vol. 40(2). – P. 239—259.

[41] Хромов А. Б. Пятифакторный опросник личности (5PFQ): учебно-методическое пособие. – Курган : КГУ, 2000. – С. 23. – ISBN: 5-86328-381-5.

[42] Costa Paul, Mccrae R. Neo PI-R professional manual // Psychological Assessment Resources. – 1992. – Vol. 396.

[43] Schwartz Shalom, Cieciuch Jan, Vecchione Michele, Davidov Eldad. Refining the Theory of Basic Individual Values // Journal of Personality and Social Psychology. – 2012. – Vol. 103. – P. 663–88.

[44] Карандашев В. Н. Методика Шварца для изучения ценностей личности: концепция и методическое руководство. Практикум по психодиагностике. – Санкт-Петербург : Речь, 2004. – 69 с. – ISBN: 5-9268-0299-7.

[45] Spyromitros-Xioufis Eleftherios, Groves William, Tsoumakas Grigorios, Vlahavas I. Multi-Label Classification Methods for Multi-Target Regression // arXiv. – 2012.

[46] Zhang Min-Ling, Zhou Zhi-Hua. A Review On Multi-Label Learning Algorithms // Knowledge and Data Engineering, IEEE Transactions on. – 2014. – Vol. 26. – P. 1819–1837.

[47] Burges Christopher. From RankNet to LambdaRank to LambdaMART: An Overview // Learning. – 2010. – Vol. 11. – URL: https://api.semanticscholar.org/CorpusID:397316.

[48] Sofiane Touati, Radjef Mohammed Said, Sais Lakhdar. A Bayesian Monte Carlo Method for Computing the Shapley Value: Application to Weighted Voting and Bin Packing Games // Computers and Operations Research. – 2020.

[49] You Gui-Rong, Shiue Yeou-Ren, Yeh Wei-Chang et al. A Weighted Ensemble Learning Algorithm Based on Diversity Using a Novel Particle Swarm Optimization Approach // Algorithms. – 2020. – Vol. 13. – P. 255.

[50] Buuren Stef. Flexible Imputation of Missing Data, Second Edition. – 2nd edition. – Chapman and Hall / CRC, 2018. – ISBN: 9780429492259.

Приложение А. Описание психометрических тестов

Таблица 15. Психометрические тесты: описательная статистика

Опросник	Признак	Код	N	Mean (SD)	Median (IQR)	Min	Max
16-факторный опросник Кеттелла	Открытость – Замкнутость	CT_1	993	9,91 (3,67)	10 (7–12)	0	19
	Эмоцион. стабильность – Неустойчивость	CT_2	993	12,99 (4,73)	13 (10–16)	0	26
	Независимость – Податливость	CT_3	993	12,84 (4,06)	13 (10–16)	1	25
	Беспечность – Озабоченность	CT_4	993	12,25 (4,35)	12 (9–15)	2	25
	Сознательность – Беспринципность	CT_5	993	10,61 (3,56)	11 (8–13)	1	20
	Смелость – Застенчивость	CT_6	993	10,88 (5,93)	11 (6–15)	0	26
	Чувственность – Твердость	CT_7	993	11,99 (3,71)	12 (10–15)	1	20
	Подозрительность – Доверчивость	CT_8	993	10,72 (3,55)	11 (8–13)	0	20
	Мечтательность – Практичность	CT_9	993	10,84 (3,02)	11 (9–13)	2	20
	Утонченность – Простота	CT_10	993	10,06 (2,99)	10 (8–12)	2	20
	Склонность к чувству вины – Спокойная самоуверенность	CT_11	993	13,99 (4,99)	14 (10–18)	0	26
	Радикализм – Консерватизм	CT_12	993	10,32 (2,97)	10 (8–12)	0	20
	Самостоятельность – Зависимость от группы	CT_13	993	12,60 (3,51)	13 (10–15)	1	20
	Сильная воля – Недостаток самоконтроля	CT_14	993	11,42 (3,43)	12 (9–14)	1	20
	Внутренняя напряженность – Расслабленность	CT_15	993	14,46 (5,25)	15 (11–18)	0	26
	Развитое мышление – Ограниченное мышление	CT_16	993	7,69 (2,32)	8 (6–9)	1	13

Продолжение на следующей странице

Таблица 15. Психометрические тесты: описательная статистика (продолжение)

Опросник	Признак	Код	N	Mean (SD)	Median (IQR)	Min	Max
Личностный	Экстраверсия	EY_1	1200	11,2 (5,38)	11 (7–15)	0	24
опросник	Психотизм	EY_2	1200	6,29 (3,31)	6 (4–8)	0	22
Айзенка	Нейротизм	EY_3	1200	16,25 (5,74)	17 (12–21)	1	25
	Искренность	EY_4	1200	10,79 (4,36)	11 (8–14)	0	25
Опросник	Гипертимность	LN_1	998	12,84 (6,47)	12 (9–18)	0	24
Леонгарда-	Дистимность	LN_2	998	14,95 (4,07)	16 (12–18)	2	24
Шмишека	Циклотимность	LN_3	998	14,71 (5,19)	15 (12–18)	0	24
	Неуравновешенность	LN_4	998	12,27 (4,83)	12 (8–16)	0	24
	Застревание	LN_5	998	11,77 (6,03)	12 (6–15)	0	24
	Эмотивность	LN_6	998	14,93 (5,85)	15 (9–18)	0	24
	Экзальтированность	LN_7	998	13,92 (4,6)	14 (10–18)	2	24
	Тревожность	LN_8	998	13,54 (5,53)	15 (9–18)	0	24
	Педантичность	LN_9	998	13,10 (4,72)	12 (9–15)	0	24
	Демонстративность	LN_10	998	15,26 (5,96)	18 (12–18)	0	24
Пятифакторный	Экстраверсия – интроверсия	BF_1	891	43,54 (11,51)	43 (35–51)	16	75
опросник	Привязанность – обособленность	BF_2	891	49,76 (11,78)	50 (42–58)	15	75
личности	Самоконтроль – импульсивность	BF_3	891	51,38 (11,34)	51 (43–60)	15	75
	Эмоциональная устойчивость – эмоциональная неустойчивость	BF_4	891	52,58 (13,96)	54 (44–63,5)	15	75
	Экспрессивность – практичность	BF_5	891	54,63 (8,68)	55 (49–61)	15	75
Ценностный	Универсализм – НИ	SC_1	747	38,66 (10)	40 (33–46)	-8	56
опросник	Безопасность – НИ	SC_2	747	25,12 (6,24)	26 (22–29)	-5	35
Шварца	Власть – НИ	SC_3	747	15,69 (6,58)	16 (11–20)	-2	28

Продолжение на следующей странице

Таблица 15. Психометрические тесты: описательная статистика (продолжение)

Опросник	Признак	Код	N	Mean (SD)	Median (IQR)	Min	Max
	Гедонизм – НИ	SC_4	747	14,61 (4,8)	15 (12–18)	-2	21
	Самостоятельность – НИ	SC_5	747	26,65 (5,35)	27 (24–30)	0	35
	Стимуляция – НИ	SC_6	747	11,36 (5,23)	12 (8–15)	-3	21
	Конформность – НИ	SC_7	747	17,04 (5,98)	18 (14–21)	-4	28
	Традиция – НИ	SC_8	747	17,36 (8,16)	18 (12–23)	-5	35
	Доброта – НИ	SC_9	747	24,06 (6,98)	25 (20–29)	-3	35
	Достижение – НИ	SC_10	747	19,49 (5,53)	20 (16–24)	-3	28
	Самостоятельность – ИП	SC_11	747	11,15 (3,44)	12 (9–14)	-1	16
	Власть – ИП	SC_12	747	4,63 (3,8)	4 (2–8)	-3	12
	Универсализм – ИП	SC_13	747	13,87 (5,8)	14 (10–19)	-6	24
	Достижение – ИП	SC_14	747	8,61 (4,54)	9 (5–12)	-4	16
	Безопасность – ИП	SC_15	747	10,69 (5,03)	11 (7–14)	-5	20
	Стимуляция – ИП	SC_16	747	5,31 (3,49)	5 (3–8)	-3	12
	Конформность – ИП	SC_17	747	6,43 (4,46)	7 (3–10)	-4	16
	Традиция – ИП	SC_18	747	4,46 (4,44)	4 (1–7)	-4	16
	Гедонизм – ИП	SC_19	747	7,38 (3,35)	8 (5–10)	-3	12
	Доброта – ИП	SC_20	747	8,27 (4,37)	9 (5–11)	-4	16
Тест Голланда	Реалистический (R)	HL_1	1278	6,42 (2,33)	6 (5–8)	0	14
	Исследовательский (I)	HL_2	1278	7,19 (2,18)	7 (6–9)	2	13
	Артистический (A)	HL_3	1278	7,09 (2,07)	7 (6–9)	1	13
	Социальный (S)	HL_4	1278	6,67 (2,04)	7 (5–8)	1	13
	Предприимчивый (E)	HL_5	1278	7,04 (2,34)	7 (5–9)	1	13
	Традиционный (C)	HL_6	1278	7,59 (2,78)	8 (6–10)	0	14

Приложение В. Интерфейс прототипа приложения

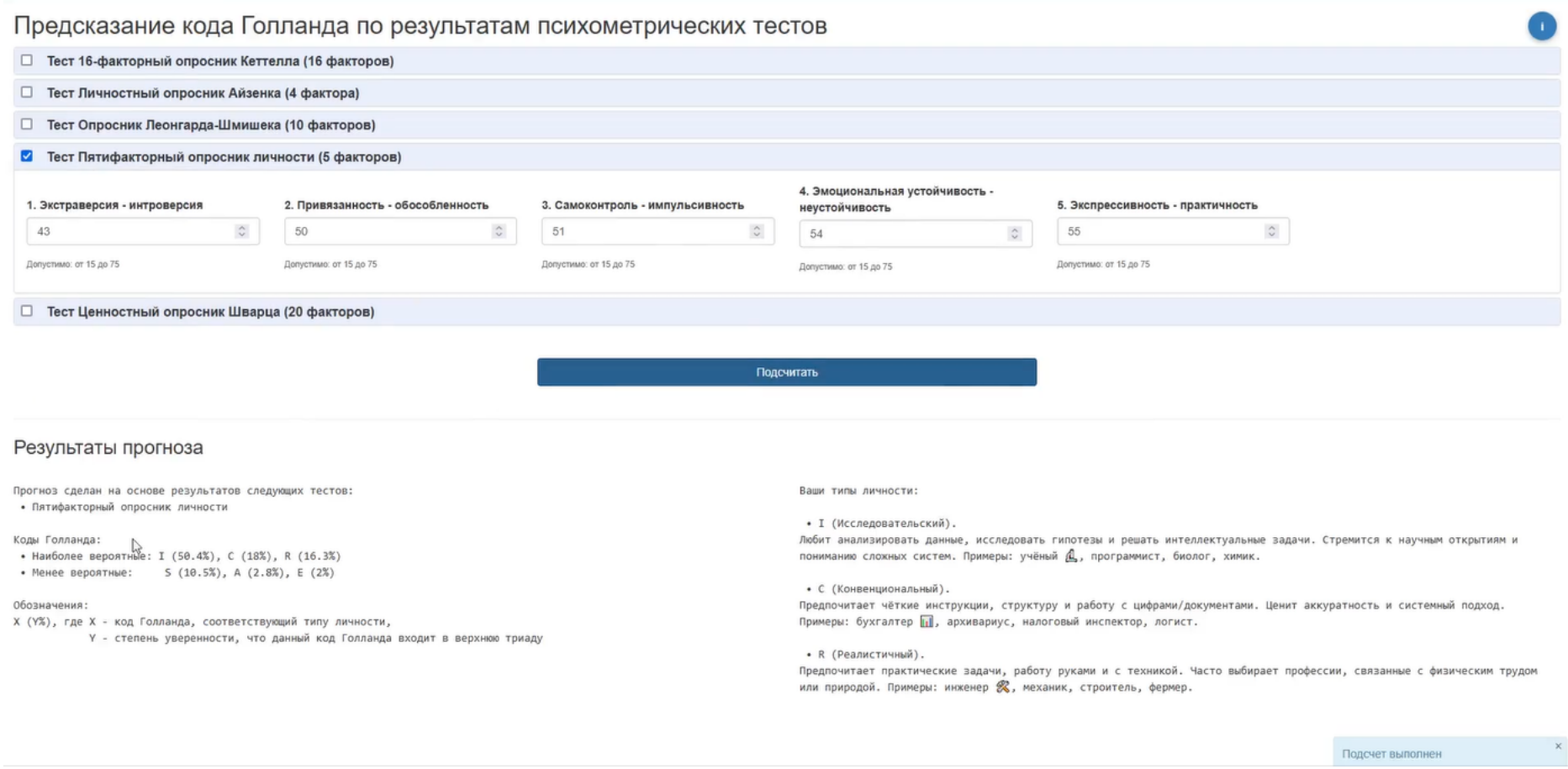

Рисунок 11. Интерфейс прототипа инструмента профориентации