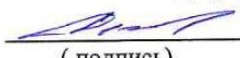


Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»
Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования

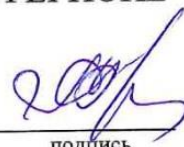
ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Директор ШПиАО
 Д.В. Денисов
(подпись) (Ф.И.О.)
« 03 » июня 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРОЕКТИРОВАНИЕ СИСТЕМЫ АНАЛИЗА ДАННЫХ ДЛЯ ОТБОРА
ЗАЯВОК НА ФИНАНСИРОВАНИЕ ЦИФРОВЫХ ПРОЕКТОВ В РЕГИОНЕ

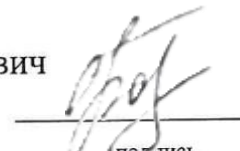
Научный руководитель: Коломыцева Анна Олеговна
к.э.н., доцент


подпись

Нормоконтролер: Огуренко Егор Владимирович


подпись

Студент группы: РИМ-220962 Борисенко Станислав Алексеевич


подпись

Екатеринбург
2024

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего
образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ
Школа профессионального и академического образования
Направление подготовки 09.04.01 Информатика и вычислительная техника
Образовательная программа 09.04.01/33.03 Инженерия машинного обучения

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студента Борисенко Станислава Алексеевича группы РИМ-
220962

(фамилия, имя, отчество)

1. Тема выпускной квалификационной работы

Проектирование системы анализа данных для отбора заявок на финансирование
цифровых проектов в регионе

Утверждена распоряжением по институту от «4» декабря 2023 г. № 33.02-05/298

2. Научный руководитель

Коломыцева Анна Олеговна, доцент кафедры ИТИСУ, кандидат экономических наук,
доцент

(Ф.И.О., должность, ученая степень, ученое звание)

3. Исходные данные к работе

Научные статьи, открытые источники

4. Перечень демонстрационных материалов

Презентация

5. Календарный план

№ п/п	Наименование этапов выполнения работы	Срок выполнения этапов работы	Отметка о выполнении
1.	<i>1 раздел (глава)</i>	до 23.03.2024 г.	<i>выполнено</i>
2.	<i>2 раздел (глава)</i>	до 29.04.2024 г.	<i>выполнено</i>
3.	<i>3–4 раздел (глава)</i>	до 20.05.2024 г.	<i>выполнено</i>
4.	<i>ВКР в целом</i>	до 24.05.2024 г.	<i>выполнено</i>

Научный руководитель Коломыцева Анна Олеговна

Ф.И.О.

Студент задание принял к исполнению 4.12.23
дата

(подпись)

(подпись)

6. Допустить Борисенко Станислава Алексеевича к защите выпускной квалификационной
работы в экзаменационной комиссии

Директор ШПиАО

(подпись)

Д.В. Денисов

Ф.И.О.

РЕФЕРАТ

Выпускная квалификационная работа магистра 56 с., 12 рис., 40 источников.

АНАЛИЗ ДАННЫХ, ПРОЕКТНОЕ ФИНАНСИРОВАНИЕ, ЦИФРОВЫЕ ПРОЕКТЫ, ИМПОРТОЗАМЕЩЕНИЕ, КРИТИЧЕСКИ ВАЖНАЯ ИНФРАСТРУКТУРА, ПРОЕКТИРОВАНИЕ СИСТЕМЫ, РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Цель работы – разработка информационной системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе.

Объект исследования – процесс отбора заявок на финансирование цифровых проектов в регионе.

Методы исследования: метод дерева решений, метод нечеткой кластеризации С-средних, экспертных оценок.

Результаты работы: разработана система анализа данных для отбора заявок на финансирование цифровых проектов в регионе.

Выпускная квалификационная работа выполнена в текстовом редакторе Microsoft Word и представлена в твердой копии.

СОДЕРЖАНИЕ

РЕФЕРАТ	1
СОДЕРЖАНИЕ	4
ВВЕДЕНИЕ	6
1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОЕКТИРОВАНИЯ СИСТЕМ АНАЛИЗА ДАННЫХ	9
1.1 Анализ существующих методов и моделей для разработки системы анализа данных, и классификация алгоритмов машинного обучения для решения задач в сфере проектного финансирования	9
1.1.1 Анализ методов и моделей для разработки системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе	9
1.1.2 Анализ предметной области, классификация алгоритмов машинного обучения	12
1.2 Описание специфики организации проекта внедрения информационной системы на основе машинного обучения	15
Выводы	16
2 МЕТОДОЛОГИЯ ПРОЕКТИРОВАНИЯ СИСТЕМЫ АНАЛИЗА ДАННЫХ ДЛЯ ОТБОРА ЗАЯВОК НА ФИНАНСИРОВАНИЕ ЦИФРОВЫХ ПРОЕКТОВ В РЕГИОНЕ	18
2.1 Подготовительный этап исследования	18
2.1.1 Формализация основных математических и аналитических зависимостей для модели машинного обучения	18
2.1.2 Методика оценки инновационных ИТ-проектов	19
2.1.3 Сбор и подготовка набора данных	21
2.2 Процессные модели этапов разработки	21
2.2.1 Процесс сбора данных	22
2.2.2 Процесс обработки данных	23
2.2.3 Процесс ручной оценки данных	25
2.2.4 Процесс разработки модели машинного обучения	26
2.2.5 Процесс обучения модели машинного обучения	28
2.2.6 Процесс преобразования модели машинного обучения и представление результатов оценки	29
2.2.7 Завершение разработки модели машинного обучения	30

2.3 Программно-аппаратный комплекс, с выделением всех ключевых преимуществ программных средств	31
2.3.1 Описание программно-аппаратного комплекса.....	31
2.3.2 Перечень программных средств	32
2.3.3 Преимущества программных средств	33
Выводы.....	34
3 РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ.....	36
3.1 Разработка прототипа системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе.....	36
3.1.1 Данные.....	36
3.1.2 Оценка набора данных.....	37
3.1.3 Создание и обучение модели машинного обучения.....	38
3.1.4 Преобразование модели машинного обучения и уменьшение пространственной размерности данных	38
3.2 Тестирование предлагаемой системы	41
3.3 Экономическая оценка перспектив внедрения прототипа системы ...	44
Выводы.....	48
ЗАКЛЮЧЕНИЕ	49
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	51

ВВЕДЕНИЕ

В современном мире информационные технологии играют ключевую роль в развитии экономики и повышении конкурентоспособности регионов. Цифровые проекты, направленные на улучшение инфраструктуры, образования, здравоохранения и других сфер жизни, требуют значительных финансовых вложений. Однако процесс отбора заявок на финансирование таких проектов часто бывает сложным и длительным, что снижает эффективность использования бюджетных средств.

Для решения этой проблемы необходимо разработать информационную систему анализа данных, которая позволит автоматизировать процесс отбора заявок, учитывая различные критерии и факторы риска. Такая система должна быть основана на процессных моделях, которые описывают последовательность и взаимосвязь этапов разработки и внедрения информационной системы. Также, для данной задачи отлично подойдет модель машинного обучения, что позволит системе определить взаимосвязь между параметрами проектов, для принятия решения о финансировании.

Вектор развития страны легко прослеживается по актуальным национальным проектам, реализация которых происходит в данный момент и является основой развития, а также намеченные правительством цели.

Одним из ключевых национальных проектов в настоящий момент является национальный проект «Цифровая экономика» [1], который включает в себя такие вопросы, как:

- регулирование цифровой среды;
- информационная инфраструктура;
- кадры для цифровой экономики;
- информационная безопасность;
- цифровые технологии;
- цифровое государственное управление;

- искусственный интеллект;
- развитие центра технологического лидерства.

Реализация цифровых проектов – это основополагающий элемент реализации национального проекта [1], что особенно остро актуализирует вопрос отбора потенциально интересных и полезных проектов, продвижение которых будет совпадать с общим вектором развития региона, а финансовая сторона не ввергнет экономику в сторону стагнации и кризиса, вместо этого возглавив всесторонний рост.

Актуальность работы обусловлена необходимостью упрощения и автоматизации процесса отбора заявок на финансирование цифровых проектов, учитывая специфику отрасли, актуальность для регионального развития, а также финансовую сторону вопроса, с учетом текущей внешнеполитической обстановки и акцентом на импортозамещение программного обеспечения, относящемуся к критически важной инфраструктуре региона. Кроме того, разработка отечественных программных средств, с целью развития региона и цифровой экономики, позволяет обеспечить реализацию национального проекта «Цифровая экономика» [1].

Объект исследования – процесс отбора заявок на финансирование цифровых проектов в регионе.

Предмет исследования – разработка системы анализа данных для эффективного отбора заявок на финансирование цифровых проектов в регионе.

Целью, данной выпускной квалификационной работы является разработка информационной системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- а) Изучение основ проектирования систем анализа данных.

- б) Выбор методов и алгоритмов анализа данных, подходящих для отбора заявок на финансирование цифровых проектов.
- в) Построение процессных моделей для реализации информационной системы.
- г) Реализация системы анализа данных и тестирование ее функциональности.
- д) Оценка эффективности разработанной системы анализа данных и предоставление рекомендаций по ее улучшению.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПРОЕКТИРОВАНИЯ СИСТЕМ АНАЛИЗА ДАННЫХ

1.1 Анализ существующих методов и моделей для разработки системы анализа данных, и классификация алгоритмов машинного обучения для решения задач в сфере проектного финансирования

1.1.1 Анализ методов и моделей для разработки системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе

Решение задачи анализа данных для отбора заявок на финансирование цифровых проектов в регионе относится к вопросам проектного финансирования.

Согласно [2-4], финансирование проектов сопряжено с высоким уровнем рисков на каждом из этапов реализации. Изначально, даже сама идея может быть неудачной и заведомо ошибочной, из-за чего проект будет обречен на закрытие. Кроме того, неправильный подход к организации и реализации может оказаться нереализуемым (например, из-за недостаточной компетентности, была допущена ошибка в планировании), что повлечет за собой как финансовые потери, так и временные. Потому крайне важно среди массы проектов, выбрать именно перспективные, корректно спланированные и в целом, актуальные проекты.

В основном, при выборе проектов, акцент делается на финансовой составляющей, сроках окупаемости и потенциальной прибыли. Безусловно, финансовую сторону вопроса нельзя игнорировать, но и учитывать исключительно только ее – также в корне неверный подход.

В рамках регионального развития и осуществления цифровой трансформации в рамках национального проекта [1], финансовая сторона вопроса отходит на второй план, уступая первенство изменениям в производственных и социальных процессах, а также проблемам, которые описаны в [5-7].

Можно выделить следующие типы методов, которые будут рассмотрены:

- методы оценки;
- методы анализа.

Основными методами оценки проектов являются [8-10]:

- экспертная оценка;
- PERT (Program Evaluation Review Technique);
- иерархия задач;
- покер планирования;
- голосование по точкам.

Метод экспертной оценки – качественный метод, который заключается в привлечении эксперта, который разбирается в вопросе и уже занимался подобного рода проектами. Метод разделяется на несколько подвидов – индивидуальная экспертная оценка и коллективную. При индивидуальной приглашает один эксперт “со стороны”. Коллективная заключается в организации некоторого совещания, на которое приглашаются непосредственно участники проекта и каждый, со стороны своей области, высказывается и дает оценку, после того как все из участников выскажутся, руководитель собирает и систематизирует результаты совещания, на основании которых подготавливается регламент действий или принимается иное решение в отношении проекта.

Метод PERT (Program Evaluation Review Technique) – количественный метод, который позволяет оценить время на реализацию конкретных задач, составных частей проекта. Такого рода оценка позволит определить обобщенное количество необходимого времени. Для большей конкретики и детализации данный метод не подходит, так как результат в лучшем случае будет неточным, а в худшем будет ошибочным и введет в заблуждение.

Метод иерархии задач – позволяет установить очередность выполнения задач проекта. Для этого используются цветные карточки, доска и участники проекта. Задачи расписываются на карточках и распределяются на доске

случайным образом. Каждый из участников поочередно подходит к доске, и может переместить любую карточку туда, где, по его мнению, она уместнее. Каждое действие должно сопровождаться комментариями и обоснованиями со стороны участника. Таким образом выстраивается очередность выполнения проекта.

Метод покер планирования – позволяет оценить сложность каждой из задачи, что достигается путем “голосования” участниками с применением специальных колод, которые состоят из последовательности Фибоначчи, знака вопроса и карточки с изображением кофе. Карточка с числом означает оценку участника, знак вопроса означает необходимость уточнения задачи, а чашка кофе – перерыв. Задачи обсуждаются последовательно. Этап обсуждения состоит из выставления каждым участником карточки “в закрытую” и открываются они лишь когда все “проголосуют”. Участники выставившие самые маленькие и самые большие оценки комментируют свое решение, а сложность задачи – сумма всех оценок. Основной недостаток – продолжительность “голосований”.

Голосование по точкам – позволяет довольно быстро оценить небольшие проекты. В данном методе применяются карточки и фишки, которые называются точками. Задачи проекта выписываются на карточки и размещаются на доске. Каждый из участников проекта получает одинаковое количество точек, чтобы таким образом оценить каждую из задач, распределив свои точки между ними. Задачи оцениваются по суммарному количеству точек, чем больше точек, тем сложнее задача.

Основными методами анализа являются:

- метод Парето;
- метод Борда;
- метод БОФа.

Рассматривая статью [11-14], наглядно демонстрируется применяемость данных методов по отношению к проектному финансированию регионального уровня, что в современных реалиях более

чем актуально в рамках цифровой трансформации в частности и реализации национального проекта [1] в целом.

Метод Парето состоит из выбора такого варианта, который по всем критериям будет не хуже первого и хотя-бы по одному будет лучше.

Метод Борда или же бальная система, преобразует критерии кандидатов в бальную систему, что в некотором роде является уравниванием показателей путем уменьшения размерности показателей к установленной балльной системе. Общепринятой является пятибалльная система, но не исключено использование любой иной системы.

Метод БОФа или метод многомерной оптимизации заключается в присваивании каждому из критериев коэффициента важности, который в совокупности с показателем критерия дает итоговую оценку. На основании итоговых оценок каждого их критериев и осуществляется выбор наиболее “весомого” варианта.

1.1.2 Анализ предметной области, классификация алгоритмов машинного обучения

В современном мире актуальны вопросы развития цифровых технологий. Все страны мира, в той или иной мере, стараются планомерно развивать как применяемые технологии, так и в целом заменять устаревшие способы взаимодействия граждан в повседневной жизни. Разумеется, столицы развиваются значительно быстрее и активнее, нежели регионы, что в особой мере выделяет вопрос развития и накладывает более строгие требования к проектам.

Вопрос финансирования проектов в регионах – это весьма сложный и трудоемкий процесс отбора претендентов, так как нет средств на всевозможные эксперименты и авантюрные проекты. Потенциальный проект, должен не только быть выгодным и рентабельным с финансовой точки зрения, но и должен быть полезным как для жителей региона, так и для тех

отраслей, которые в регионе развиты. Подобное, несколько обособленное и выделяющееся требование позволит не распылять средства, которых и так не слишком много, и при этом развить уже устоявшиеся отрасли.

Одной из главных проблем [15-16] является недостаток финансирования. Многие проекты требуют значительных инвестиций, которые не всегда доступны для региональных властей. Кроме того, банки часто не готовы предоставлять кредиты на длительные сроки и под низкие проценты, что также затрудняет реализацию проектов.

Еще одной проблемой является отсутствие опыта у региональных властей в управлении проектами такого масштаба. Часто руководители регионов не имеют достаточной квалификации и опыта для эффективного управления проектами.

Кроме того, существуют проблемы с качеством проектной документации и оценкой рисков. Не всегда проекты имеют полную и точную информацию о затратах, сроках и возможных рисках, что может привести к неудаче проекта.

Для решения этих проблем необходимо разрабатывать новые механизмы финансирования, такие как государственно-частное партнерство, привлекать инвестиции из различных источников, а также обучать руководителей регионов управлению проектами.

Реализация цифровых проектов позволяет развивать цифровую экономику страны, реализуя тем самым указы президента и национальные проекты страны, которые направлены на всестороннее развитие применяемых технологий. Также, таким образом происходит цифровая трансформация технологического развития и осуществляется расширение цифрового потенциала.

Существует множество алгоритмов машинного обучения, которые могут быть использованы для решения различных задач в сфере проектного финансирования. Некоторые из них включают в себя [17-20]:

- деревья решений;

- нейронные сети;
- регрессионный анализ;
- кластеризация;
- факторный анализ;
- анализ главных компонент;
- байесовский анализ.

Деревья решений являются одним из самых популярных алгоритмов машинного обучения. Они позволяют классифицировать данные и определять наиболее важные факторы, влияющие на успешность проектного финансирования. Преимуществом деревьев решений является их способность работать с большими объемами данных и определять сложные зависимости между входными параметрами и результатом. Однако, деревья решений могут быть подвержены переобучению и требуют тщательной предварительной обработки данных.

Нейронные сети также широко используются в проектном финансировании. Они позволяют моделировать сложные зависимости между входными данными и результатом, а также могут обрабатывать большие объемы данных. Однако, нейронные сети требуют большого количества данных для обучения и могут быть сложными в настройке.

Регрессионный анализ используется для прогнозирования будущих значений на основе исторических данных. Этот метод позволяет оценить влияние различных факторов на успешность проекта и предсказать его результаты. Однако, регрессионный анализ требует большого количества данных и может быть подвержен проблемам мультиколлинеарности.

Кластеризация используется для разделения данных на группы на основе их сходства. Этот метод может помочь выявить закономерности в данных и определить наиболее перспективные проекты. Однако, кластеризация может быть сложной в интерпретации и требует дополнительных методов анализа.

Факторный анализ используется для выявления скрытых закономерностей в данных. Он позволяет сократить размерность данных и получить более компактное представление информации. Однако, факторный анализ может быть сложным в интерпретации и требует знаний в области статистики.

Анализ главных компонент используется для уменьшения размерности данных без потери информации. Этот метод помогает выявить наиболее значимые факторы, влияющие на успех проекта, и может быть использован для снижения затрат на анализ данных. Однако, анализ главных компонент может быть сложным для интерпретации и требует дополнительной обработки данных.

Байесовский анализ используется для оценки вероятности события на основе имеющихся данных. Этот метод позволяет получить более точные результаты, чем классический статистический анализ. Однако, байесовский анализ требует знания теории вероятностей и может быть сложным для интерпретации.

Решение задач в сфере проектного финансирования достигается путем анализа данных и применения различных методов. Машинное обучение позволяет автоматизировать этот процесс, а также помогает в принятии решений о финансировании проектов.

1.2 Описание специфики организации проекта внедрения информационной системы на основе машинного обучения

Специфика организации проекта внедрения информационной системы на основе машинного обучения (ML) в сфере проектного финансирования является сложной и многогранной задачей. Она включает в себя несколько этапов, каждый из которых имеет свои особенности [21-23]:

- анализ данных;
- разработка алгоритма;

- обеспечение безопасности данных;
- обучение системы;
- интеграция системы.

Первым этапом является анализ данных. На этом этапе необходимо определить, какие данные необходимы для принятия решения о финансировании проекта, и как эти данные должны быть обработаны. Для этого могут использоваться различные методы машинного обучения, такие как регрессионный анализ, кластеризация, нейронные сети и другие.

Второй этап – разработка алгоритма. Алгоритм должен обрабатывать данные и на основе их анализа принимать решение о финансировании проекта. Этот алгоритм может быть разработан с использованием различных языков программирования, таких как Python, R или Java.

Третий этап – обеспечение безопасности данных. Данные, используемые для принятия решений о финансировании проектов, могут содержать конфиденциальную информацию, такую как персональные данные или коммерческая тайна. Поэтому необходимо обеспечить защиту этих данных от несанкционированного доступа.

Четвертый этап – обучение системы. После разработки алгоритма необходимо обучить систему на имеющихся данных и проверить ее работу. Это может включать в себя тестирование алгоритма на различных наборах данных и оценку его точности и эффективности.

Пятый этап – интеграция системы с другими информационными системами. В основном заключается в разработке механизмов обмена данными, протоколов передачи данных и обеспечения безопасности передачи данных.

Выводы

В ходе данной главы, были изучены теоретические основы проектирования систем анализа данных. Данный вопрос включает в себя

анализ методов и моделей, которые будут в дальнейшем использованы при разработке модели машинного обучения, что позволяет определиться с вектором развития проекта и позволяет обрести конкретику в задаче проектирования и последующей разработке. Кроме того, был проведен анализ предметной области, проведена классификация алгоритмов машинного обучения.

В рамках изучения теоретических основ, была описана специфика организации проекта внедрения информационной системы, основанной на машинном обучении. Были выделены следующие этапы реализации проекта:

- анализ данных;
- разработка алгоритма;
- обеспечение безопасности данных;
- обучение системы;
- интеграция системы.

Таким образом, была изучена теория, что позволит продолжить работу, переходя непосредственно к процессу проектирования системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе.

2 МЕТОДОЛОГИЯ ПРОЕКТИРОВАНИЯ СИСТЕМЫ АНАЛИЗА ДАННЫХ ДЛЯ ОТБОРА ЗАЯВОК НА ФИНАНСИРОВАНИЕ ЦИФРОВЫХ ПРОЕКТОВ В РЕГИОНЕ

2.1 Подготовительный этап исследования

2.1.1 Формализация основных математических и аналитических зависимостей для модели машинного обучения

В современном мире машинное обучение стало неотъемлемой частью многих сфер деятельности, таких как обработка данных, анализ информации и прогнозирование. Однако для успешного применения методов машинного обучения необходимо глубокое понимание математических основ, лежащих в основе алгоритмов.

Математика в машинном обучении включает в себя такие разделы, как линейная алгебра, теория вероятностей и многомерное исчисление. Каждый из этих разделов играет свою роль в процессе машинного обучения и необходим для создания эффективных и точных моделей.

Линейная алгебра является основой для представления систем линейных уравнений и используется для решения задач оптимизации. Теория вероятностей позволяет оценить степень достоверности гипотез и предположений, а многомерное исчисление применяется для численной оптимизации алгоритмов машинного обучения.

Важно отметить, что понимание математики в машинном обучении не ограничивается только теоретическими знаниями [24-26]. Специалисты должны также уметь применять эти знания на практике, используя доступные вычислительные библиотеки, такие как NumPy.

Формализация математических и аналитических зависимостей имеет большое значение для машинного обучения (ML), так как помогает специалистам лучше понять алгоритмы и методы, используемые в ML. Это неотъемлемая часть роли специалиста по работе с данными.

Помимо вопроса разработки и проработки модели машинного обучения, не менее важным, а местами и более, являются вопросы сбора, подготовки и обработки данных, которые будут использоваться при модели.

Для решения задачи отбора заявок на финансирование цифровых проектов в регионе необходимо определить, что является решающим фактором и каким образом будет организовано выполнение поставленной задачи.

В современном мире для России наиболее остро стоят вопросы импортозамещения в критически важных областях, обеспечивая тем самым для страны независимость от внешнего влияния на обороноспособность и экономическое развития, в случаях неблагоприятного внешнеполитического развития событий с недружественными странами, перечень которых постоянно актуализируется правительством Российской Федерации [27].

Таким образом определим доминирующий фактор, который играет решающую роль при выборе ИТ-проектов для успешного развития региона – импортозамещение иностранного программного обеспечения, в первую очередь в критической ИТ-инфраструктуре региона.

2.1.2 Методика оценки инновационных ИТ-проектов

Прежде чем переходить к сбору и подготовке данных, необходимо определить методику оценки ИТ-проектов. Данная методика будет основанием для формирования набора данных, который послужит основой для обучения ML-модели.

Примем допущение, что жизненный цикл информационного ИТ-проекта – его жизненный цикл в венчурном фонде, т.е. от момента поступления заявки на предоставление финансирования до выхода фонда из проекта путем, например, продажи своей доли в проекте.

Этап 1. Выбор квалифицированных экспертов, принимающих непосредственное участие в первоначальной оценке инновационных ИТ-

проектов. Во время проведения данного этапа необходимо привлечь такое число экспертов, квалификация которых будет достаточна для того, чтобы принять участие в процедуре оценки. Данный этап является настолько обширным и разносторонним, что может быть выделен в качестве полноценной отдельной научной работы и выходит за рамки данной выпускной квалификационной работы. Сделаем допущение, что процесс набора и подтверждения требуемой квалификации экспертов, проводимый конкретным венчурным фондом, является валидной и позволяет собрать такую группу экспертов, которая может осуществить первоначальную оценку инновационных ИТ-проектов.

Результатом этапа является численность экспертной группы и обозначается как m .

Этап 2. Выбор и утверждение иерархии показателей ИИТП.

Этап 3. Нечеткая кластеризация ИИТП, формулирование следующих групп проектов:

- проекты, которые финансировать целесообразно – K_1 ;
- проекты, которые требуют доработки – K_2 ;
- проекты, которые финансировать не целесообразно – K_3 .

В качестве алгоритма был выбран алгоритм Густафсона-Кесселя или алгоритм нечеткой кластеризации $\Lambda_C: IP \rightarrow K$, согласно результатам различных исследований и расчетов [28-32], поскольку является одним из лучших в рамках реализации нечеткой кластеризации, т.к. позволяет находить кластеры в виде гиперэллипсоидов разной ориентации, в отличие от того-же алгоритма с-средних, который находит кластеры одинаковой сферической формы. Кроме этого, выбранный алгоритм может использоваться для поиска линейных кластеров, что невозможно для алгоритма с-средних.

2.1.3 Сбор и подготовка набора данных

Для подготовки данных, которые впоследствии будут использоваться для обучения и проверки модели, привлекается экспертная группа. Группа экспертов выполнит подбор и оценку различных инвестиционных проектов, затрагиваемая область которых напрямую связана с критически важной инфраструктурой региона.

Подбор ИТ-проектов для формирования набора данных заключается в вычлениении таких проектов, которые так или иначе могут быть задействованы в рамках глобальной инициативы по импортозамещению программного обеспечения, согласно [33]. Кроме того, должны быть представлены и заведомо неподходящие проекты, чтобы при обучении модели не допустить “заучивание” моделью идеального варианта набора данных, что впоследствии приведет к ее некорректной работе в случаях, когда на вход модели будут поданы различные проекты, которые даже идейно не соответствуют необходимому критерию причастности к критически важной инфраструктуре региона.

Оценка ИТ-проектов заключается в “ручной” обработке группой экспертов набора данных, согласно определенным интересующим критериям. Это необходимо, чтобы в дальнейшем научить модель самостоятельно оценивать проекты исключительно на основе характеристик ИТ-проектов.

2.2 Процессные модели этапов разработки

Процесс разработки информационной системы – довольно трудоемкий и сложный, по своей сути, процесс. В случае с разработкой ML-модели, помимо самого программирования и остальных базовых процессов, связанных с применением машинного обучения в проекте [34-37],

необходимо собрать и подготовить данные. Данный процесс, можно разбить на следующие этапы:

- а) Сбор данных.
- б) Обработка данных.
- в) Проведение ручной оценки данных экспертной группой.
- г) Разработка ML-модели.
- д) Обучение ML-модели.
- е) Преобразование ML-модели и представление результатов оценки.
- ж) Завершение разработки ML-модели.

Хоть этапы и пронумерованы, достижение шестого пункта совершенно не означает завершение процесса разработки и переход к седьмому пункту. Первые этапы, связанные с подготовкой данных и разработкой модели, далеко не всегда могут быть протестированы досконально, с выявлением всех ошибок и возможных неточностей. Полная и исчерпывающая проверка возможна лишь в момент обучения модели. Таким образом, достигнув этапы обучения и преобразования модели, возможно выявление ошибок и недоработок, которые относятся к предыдущим этапам, что потребует возврата к соответствующим этапам с корректировками и соответствующими изменениями. Повторять «возврат» необходимо до тех пор, пока не будут учтены и исправлены все выявленные неточности и ошибки, и пока модель не будет удовлетворять необходимому функционалу и требованиям.

2.2.1 Процесс сбора данных

Данные в моделях машинного обучения играют наиболее важную роль. Вклад данных в успешный результат разработки является решающим и наиболее ценным, ведь уже на этапе обучения модели закладывается фундамент функциональных возможностей и корректность выполнения поставленных задач.

Для начала необходимо собрать данные и сформировать набор данных. Сбором занимается привлеченная группа экспертов. Источниками информации, к которым обращаются эксперты, являются:

- открытые источники;
- архивные данные прошедших региональных тендоров;
- стартапы.

Процесс сбора данных представлен в виде процессной модели и изображен на рисунке 2.1.



Рисунок 2.1 – Процессная модель сбора данных группой экспертов

2.2.2 Процесс обработки данных

После того, как экспертная группа подготовила и сформировала набор данных, необходимо обработать полученные данные. Первоначально, данные являются «грязными», что выражается в вероятных аномалиях, ошибках при заполнении (человеческий фактор), заведомо некорректные данные (например, из-за серверной ошибки, во время записи данных). Чтобы

«очистить» данные, необходимо провести обработку, заполнить пустые поля (если пустое значение не подразумевает, и является ошибкой), используя различные подходы (среднее значение, аппроксимированное среднее значение и т.д.), также необходимо поработать над аномальными значениями, которые впоследствии повлияют на модель и не позволят достичь необходимых результатов по качеству и точности выполнения. В случае, когда данных достаточно, а количество аномальных значений невелико (по отношению к общему числу данных), допустимо удалить такого рода аномалии, в противном случае (когда данных не так уж и много, или в случаях значительного количества аномалий) необходимо проводить преобразования (например, среднее значение).

Процесс обработки набора данных представлен в виде процессной модели и изображен на рисунке 2.2.

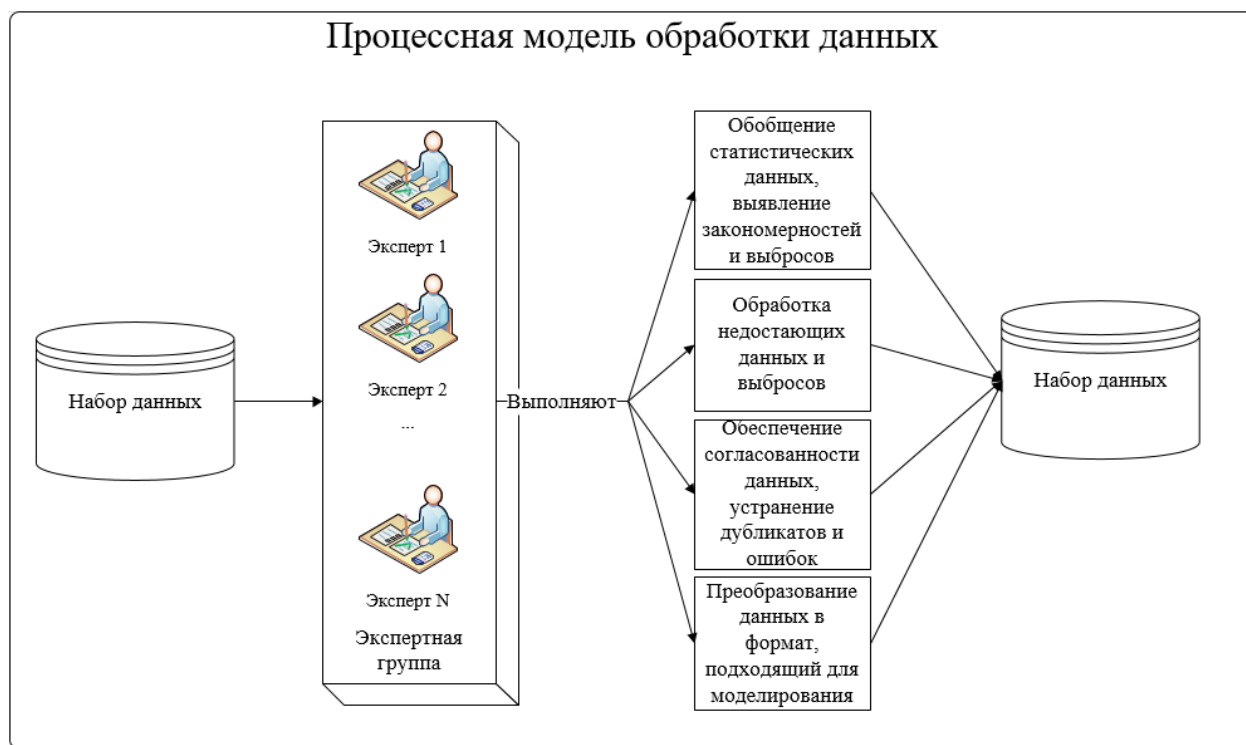


Рисунок 2.2 – Процессная модель обработки набора данных

2.2.3 Процесс ручной оценки данных

Данные собраны и обработаны, самое время провести анализ, т.е. то, чем в дальнейшем должна будет заниматься модель машинного обучения. Ранее была описана методика проведения данного анализа, согласно которой экспертная группа будет осуществлять оценку ИТ-проектов, что позволит обучить модель и в дальнейшем, использовать ее. Оценка производится по следующим параметрам:

- основные параметры ИТ-проектов;
- отношение к критически важной инфраструктуре.

Безусловно, основные параметры, которые важны при оценке ИТ-проектов важны, но ввиду текущей внешнеполитической обстановки, важно в первую очередь выделить и развивать те проекты, которые напрямую влияют на жизнеспособность, развитие и обороноспособность России. В связи с чем, оценка отношения ИТ-проекта к критически важной инфраструктуре региона является неотъемлемой и чуть-ли не основной целью оценивания, конечно же учитывая основные параметры, на основании которых строится мнение о том, насколько ИТ-проект жизнеспособен, выгоден и целесообразен ли.

Процесс ручной оценки экспертной группой представлен в виде процессной модели и изображен на рисунке 2.3.

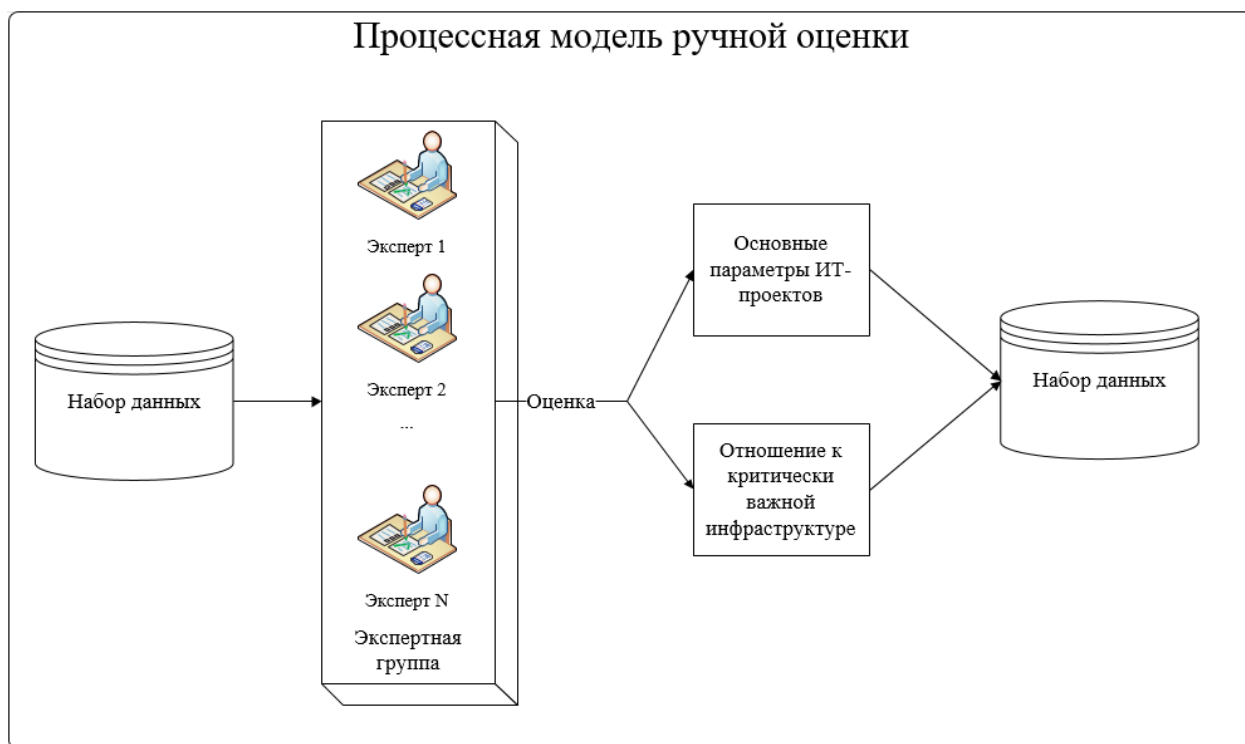


Рисунок 2.3 – Процессная модель ручной оценки набора данных ИТ-проектов группой экспертов

2.2.4 Процесс разработки модели машинного обучения

Для реализации информационной системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе могут быть использованы следующие модели машинного обучения:

а) Деревья решений (CART) – для определения потенциальных интересов поставщиков к закупкам и предоставления индивидуальных подборок закупок, которые могут их заинтересовать.

б) Ансамблевые методы – для повышения точности и скорости инференса, такие как Random Forest, Gradient Boosting и XGBoost.

в) Логистическая регрессия – для определения вероятности одобрения заявок, на основе предоставленных данных.

г) Нечеткая кластеризация – позволяет учитывать разную степень принадлежности заявок к различным кластерам, что может быть полезным при оценке и отборе проектов.

д) Метод k-ближайших соседей (kNN) – для классификации заявок на основе их сходства с другими успешными заявками.

е) Метод опорных векторов (SVM) – для разделения заявок на классы на основе предоставленных данных.

Метод нечеткой кластеризации лучше других методов подходил для реализации информационной системы анализа данных для отбора заявок на финансирование цифровых проектов регионе, потому что [ссылка на источник?]:

- определяет количество факторов риска всех данных и зависимость степени безопасности при вычислении взаимной информации;
- использует алгоритм K-means кластеризации для классификации данных и поиска оптимального значения для каждой степени риска;
- позволяет избежать проблем чувствительности к начальному значению, нелинейности и сложности оценки рисков;
- простой в реализации и легко рассчитывается.

Выбранный метод кластеризации имеет существенное преимущество перед другими методами, так как учитывает степень принадлежности к каждому кластеру. Эта особенность позволяет более точно определить принадлежность элементов к группам и улучшить качество результирующей информации.

Процесс разработки модели нечеткой кластеризации представлен в виде процессной модели и изображен на рисунке 2.4.



Рисунок 2.4 – Алгоритм разработки модели нечеткой кластеризации

2.2.5 Процесс обучения модели машинного обучения

Процесс обучения модели – становление вектора функционального назначения модели, который определит возможности, корректность, а также точность работы. Для осуществления процесса обучения, необходимо использовать подготовленный ранее экспертной группой набор данных.

Далее, модель сохраняется в нужном формате, и считается готовой к тому, чтобы быть протестированной в рамках распределения ИТ-проектов по определенным категориям (кластерам).

Процесс обучения модели представлен в виде процессной модели и изображен на рисунке 2.5.

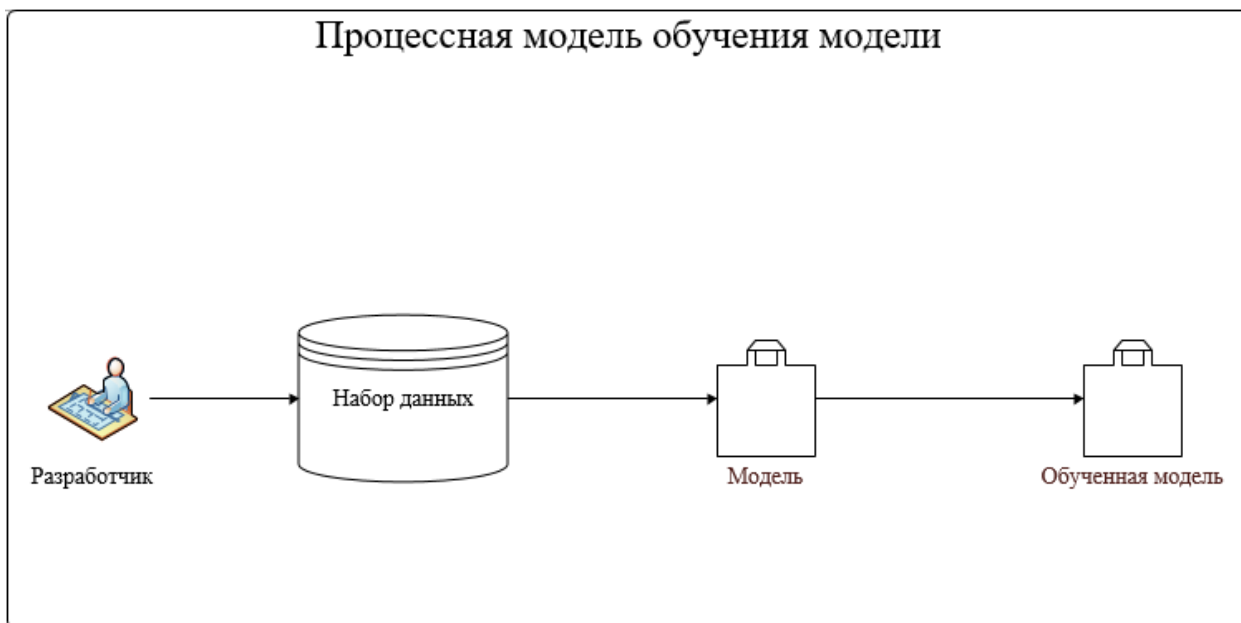


Рисунок 2.5 – Процессная модель обучения модели

2.2.6 Процесс преобразования модели машинного обучения и представление результатов оценки

Процесс преобразования модели машинного обучения и представление результатов оценки служит финальной стадией становления продукта, перед тем как модель можно будет использовать в рамках тех задач, которые были на нее возложены, а именно – отбор заявок на финансирование цифровых проектов в регионе. Конечно, финальным он будет лишь в случае успешного проведения. Если же результаты оценки не будут удовлетворять требованиям, произойдет шаг назад, к предыдущим этапам, для внесения изменений и устранению недочетов.

Процесс может завершиться неудачно в следующих случаях [38-40]:

- недостаточная точность, выявленная по результатам метрического анализа;
- ложные предпосылки, ожидание слишком высокого качества работы модели на основе модельных данных, а не реальных;
- завышенное доверие к системе, чрезмерное ожидание точных результатов без учёта возможных ухудшений;

– проблема изменений, изменение бизнес-оценки результатов работы модели, качества или природы входных данных.

Процесс преобразования модели нечеткой кластеризации и предоставление результатов оценки представлен в виде процессной модели и изображен на рисунке 2.6.



Рисунок 2.6 – Процессная модель преобразования модели

2.2.7 Завершение разработки модели машинного обучения

ML-модель считается готовой к использованию на реальных данных, когда после обучения и преобразования результаты оказались достаточными, а функциональные возможности полностью соответствуют ожидаемым. В таком случае, модель можно считать готовой к дальнейшему применению и тестированию на реальных данных, которые уже не будут иметь результата анализа, а сама модель будет заменять экспертную группу.

2.3 Программно-аппаратный комплекс, с выделением всех ключевых преимуществ программных средств

2.3.1 Описание программно-аппаратного комплекса

В рамках реализации системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе, в соответствии с логическими слоями, являющимися составляющими компонентами программно-аппаратного комплекса, выполняются следующие части:

- слой бизнес-логики;
- слой работы с данными.

В качестве реализации слоя бизнес логики выступает сама ML-модель, осуществляющая процесс анализа потенциально перспективных ИТ-проектов, которые и разделение проектов на кластеры, перечень которых был определен ранее.

Слой работы с данными состоит из формируемого и подготавливаемого набора данных экспертной группой, а также дальнейшее взаимодействие с данными во время формирования и обучения модели.

Программно-аппаратный комплекс для отбора заявок на финансирование цифровых проектов в регионе может быть основан на использовании ML-модели, которая будет анализировать данные о проектах и предоставлять рекомендации по их финансированию.

Модель должна учитывать различные факторы, такие как:

- техническая и финансовая состоятельность;
- экспансионные планы;
- бизнес-модель и монетизация;
- рыночный потенциал;
- инновационность продукта и (или) технологии;
- оценка компетентности команды;
- соответствие целям инвестора;

– степень отношения к критически важной инфраструктуре региона.

Основываясь на данных факторах, модель должна будет сделать вывод о том, к какому из кластеров отнести тот или иной ИТ-проект. Перечень распределенных проектов и будет являться результатом работы модели, иными словами это будет перечень рекомендаций, с выделением однозначно рекомендуемых, вероятно подходящих и совсем не подходящих проектов.

Для обучения модели необходимо собрать данные о успешно профинансированных и отклоненных проектах, а также провести анализ этих данных с использованием методов машинного обучения, таких как деревья решений, случайные леса или градиентный бустинг. Сбор данных отведен группе экспертов, а сам процесс более подробно был расписан ранее.

После обучения модель сможет автоматически анализировать новые заявки на финансирование и предоставлять рекомендации по их одобрению или отклонению. Это позволит региональным властям быстрее и эффективнее принимать решения о финансировании цифровых проектов, способствуя развитию цифровой экономики региона, а также, что более важно, обеспечить независимость от влияния и прочих решений недружественных для Российской Федерации стран, преследующих цели снижения обороноспособности страны и продавливания интересов, которые противоречат интересам и развитию страны.

2.3.2 Перечень программных средств

Python следует выбрать для реализации кластерной модели машинного обучения, потому что это высокоуровневый язык программирования с динамической семантикой, простой для работы и чтения. У Python отличная производительность при обработке данных, множество библиотек и фреймворков, таких как Numpy, SciPy, SciKit-Learn, TensorFlow, CNTK и

Apache Spark. Python также понятен и гибок, что делает его идеальным выбором для машинного обучения.

Для реализации системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе, используя язык программирования Python, могут быть использованы следующие пакеты:

- scikit-learn;
- numpy;
- pandas;
- scipy;
- fcmeans;
- matplotlib;
- seaborn.

Пакет «scikit-learn» не подходит для реализации алгоритмов нечеткой кластеризации, так как он предназначен для работы с четкими данными и не содержит встроенных функций для выполнения этой задачи.

Остальные пакеты, в том или ином виде, являются вторичными, помогающими проводить анализ данных, представлять данные в нужном формате, строить графики.

2.3.3 Преимущества программных средств

Основным пакетом, который будет использован для реализации метода нечеткой кластеризации, был выбран пакет «fcmeans».

Пакет «fcmeans» в языке программирования Python используется для реализации метода Fuzzy C-Means (FCM) кластеризации. FCM — это метод мягкой кластеризации, который позволяет определять центры кластеров и вероятности принадлежности точек данных к этим кластерам. В пакете «fcmeans» реализованы алгоритмы для вычисления центров кластеров и вероятностей принадлежности, а также функция для визуализации результатов кластеризации.

Ключевыми особенностями пакета «fmeans» являются:

- а) Возможность обрабатывать большие объёмы данных. FMeans эффективно обрабатывает данные с миллионами точек, что делает его подходящим для анализа больших наборов данных.
- б) Высокая точность результатов. FMeans использует метод наименьших квадратов для определения оптимального разбиения данных на кластеры, что обеспечивает высокую точность результатов.
- в) Гибкость в выборе количества кластеров. FMeans позволяет определить оптимальное количество кластеров на основе различных критериев, таких как коэффициент силуэта или индекс Дэвиса-Боулдина.
- г) Устойчивость к выбросам и шумам. FMeans использует метод наименьших квадратов для определения оптимального разбиения данных на кластеры, что делает его устойчивым к выбросам и шумам.
- д) Простота использования и интерпретации результатов. FMeans предоставляет простой и понятный интерфейс для определения количества кластеров и визуализации результатов. Результаты легко интерпретируются, что облегчает принятие решений, на основе полученных данных.

Выводы

Таким образом, был проведен подготовительный этап исследования, который включает в себя формализацию основных математических и аналитических зависимостей для модели машинного обучения, подготовку методики оценки инновационных ИТ-проектов, а также описание процесса сбора и подготовки набора данных, который в дальнейшем будет использован для обучения и тестирования модели.

В рамках формирования и описания методики оценки инновационных ИТ-проектов, были определены группы разбиение на следующие группы проектов:

- проекты, которые финансировать целесообразно;

- проекты, которые требуют доработки;
- проекты, которые финансировать не целесообразно.

Данная группировка, в дальнейшем, будет применена для реализации кластеризации проектов, определяя и количество, и определения кластеров.

Также, был определен перечень процессов, с подготовкой процессной модели для каждого. Процессы, необходимые для достижения цели данной выпускной квалификационной работы, представляют из себя следующий перечень:

- а) Сбор данных.
- б) Обработка данных.
- в) Проведение ручной оценки данных экспертной группой.
- г) Разработка ML-модели.
- д) Обучение ML-модели.
- е) Преобразование ML-модели и представление результатов оценки.
- ж) Завершение разработки ML-модели.

В рамках данной главы был определен программно-аппаратный комплекс, необходимый и достаточный для реализации информационной системы, а также были выделены преимущества программных средств, использование которых предполагается при разработке модели машинного обучения.

3 РАЗРАБОТКА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

3.1 Разработка прототипа системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе

Метод нечеткой кластеризации, выбранный для реализации ML-модели, которая предназначена для реализации информационной системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе, представляет из себя разбиение данных на кластеры, согласно обозначенным результирующим вариантам. К таким вариантам относятся:

- а) Проекты, которые финансировать целесообразно.
- б) Проекты, которые требуют доработки.
- в) Проекты, которые финансировать не целесообразно.

Реализация модели осуществляется с использованием библиотеку языка программирования Python «fcmmeans». Данная библиотека предоставляет функционал, который полностью реализует функционал метода нечеткой кластеризации, что достаточно для реализации ML-модели проектного финансирования.

3.1.1 Данные

В ходе работы экспертной группы, был подготовлен набор данных, который содержит в себе информацию о потенциальных проектах, которые были вручную оценены экспертами и представлены для дальнейшего обучения модели. Сбор и обработка данных осуществлялась согласно описанным ранее процессам, результатом чего является набор данных, описывающих потенциальные ИТ-проекты, которые также еще необходимо оценить.

Оценка проектов осуществлялась также экспертной группой, согласно описанным процессам и описанной методике. Пример получившихся

данных, что является выдержкой из набора данных, представлен на рисунке 3.1.

	A1	A2	A3	A4	A5	A6	A7	A8
Номер проекта								
1	0.2124	0.6615	0.6822	0.1251	0.1350	0.0909	0.0682	0.3932
2	0.0378	0.4860	0.1026	0.1584	0.8433	0.6741	0.0103	0.5354
3	0.2925	0.2592	0.8964	0.7587	0.1251	0.1251	0.0896	0.3775
4	0.5499	0.6750	0.0603	0.1701	0.1170	0.3744	0.0060	0.9027
5	0.5697	0.2115	0.0306	0.5427	0.0963	0.5247	0.0031	0.9007

Рисунок 3.1 – Пример данных, собранных экспертной группой

3.1.2 Оценка набора данных

Для реализации ML-модели необходимо оценить различные параметры набора данных, оценивая на наличие выбросов, аномалий, потенциально ошибочных данных, которые так или иначе могут повлиять на конечный результат функционирования модели, ее точность и корректность работы. К таким параметрам относятся: среднее значение, стандартное отклонение, минимальное значение, максимальное значение, процентиль 25%, процентиль 50% и процентиль 75%.

Согласно параметрам, набор данных был должным образом обработан и подготовлен экспертной группой, и является пригодным для дальнейшего использования в рамках обучения ML-модели. Параметры оценки набора данных представлены на рисунке 3.2.

	A1	A2	A3	A4	A5	A6	A7	A8
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	0.479444	0.371322	0.336894	0.428250	0.395358	0.387898	0.033988	0.511876
std	0.281644	0.243847	0.262873	0.294204	0.279731	0.270552	0.026007	0.297979
min	0.019800	0.013300	0.002700	0.005700	0.003700	0.001800	0.000300	0.026900
25%	0.205200	0.177525	0.085950	0.133425	0.138825	0.151650	0.008950	0.344900
50%	0.549900	0.302850	0.293400	0.403650	0.366750	0.357750	0.029350	0.465700
75%	0.742050	0.581175	0.497250	0.694575	0.650475	0.606825	0.049700	0.694950
max	0.899100	0.865800	0.896400	0.892800	0.858600	0.895500	0.089600	0.992200

Рисунок 3.2 – Анализ набора данных

3.1.3 Создание и обучение модели машинного обучения

Оценив набор данных и убедившись в его готовности, переходим к следующему этапу – формированию и последующее обучение модели.

Результатом обучения модели станет определение моделью центров кластеров, согласно переданным ей данным, на основании чего можно будет выделить то, к какому из вариантов был отнесен тот или иной ИТ-проект, определяя целесообразность реализации каждого.

Для возможности корректно представления информации и оценки работы модели, необходимо уменьшить пространственную размерность получившихся данных, специально создав для этого отдельную ML-модель.

3.1.4 Преобразование модели машинного обучения и уменьшение пространственной размерности данных

Для реализации уменьшения пространственной размерности данных, в нашем случае, создается еще одна модель, которая основана на методе главных компонент или PCA (англ. principal component analysis). Данное действие позволит уменьшить шестимерное пространство оценок и центров

кластеров до двумерного, что позволит представить результат работы, который будет понятен и нагляден для визуального представления.

Итогом преобразования станет новый набор данных, включающий в себя:

- условное расстояние относительно центров кластеров;
- метка кластера, к которому относится проект;
- условные координаты расположения проекта в двумерном пространстве.

На рисунке 3.3 представлен график распределения ИТ-проектов между кластерами.

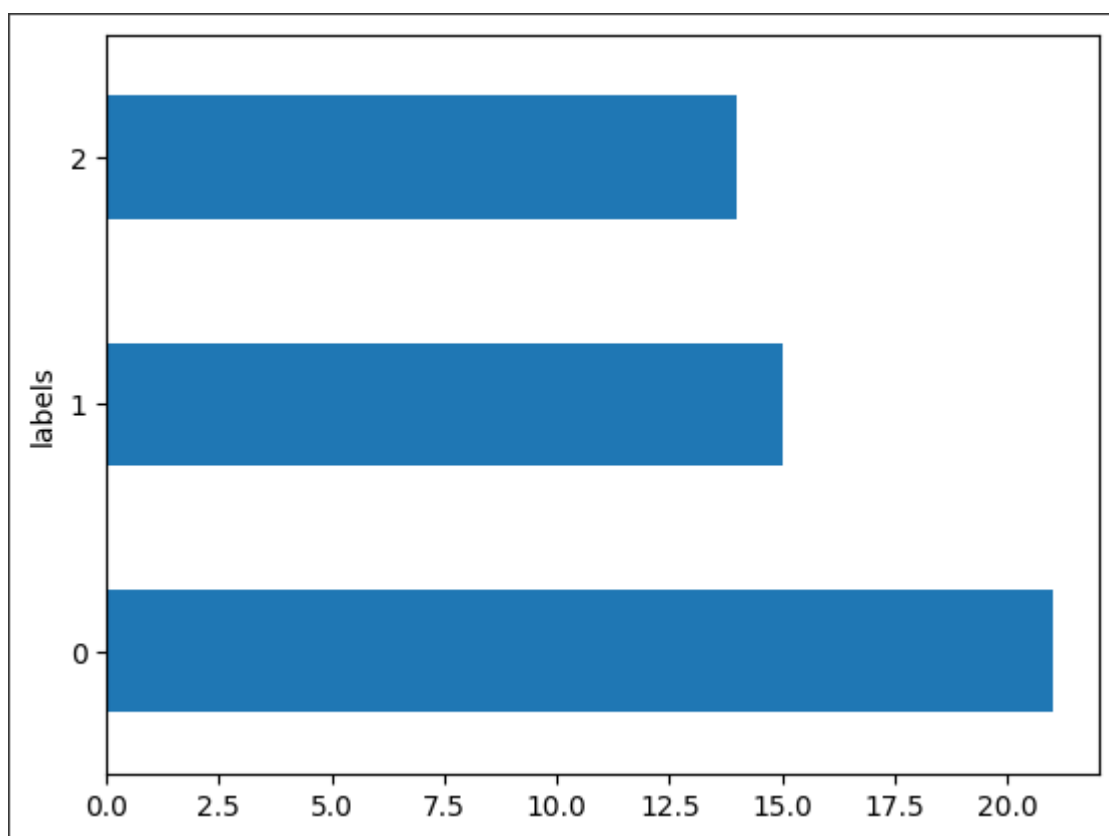


Рисунок 3.3 – График распределения ИТ-проектов между кластерами

На рисунке 3.4 представлен график распределения параметров ИТ-проектов между кластерами.

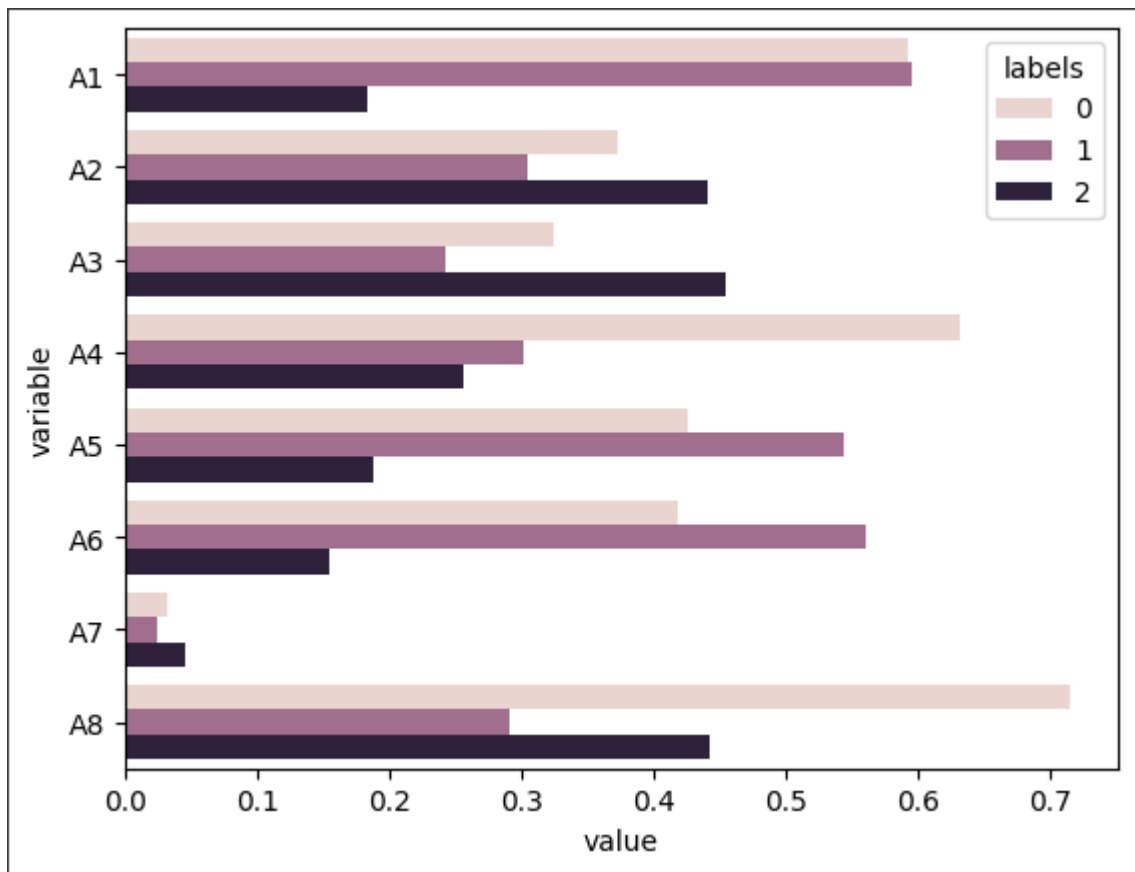


Рисунок 3.4 – График распределения параметров ИТ-проектов между кластерами

На рисунке 3.5 представлено наглядное распределение проектов по кластерам.

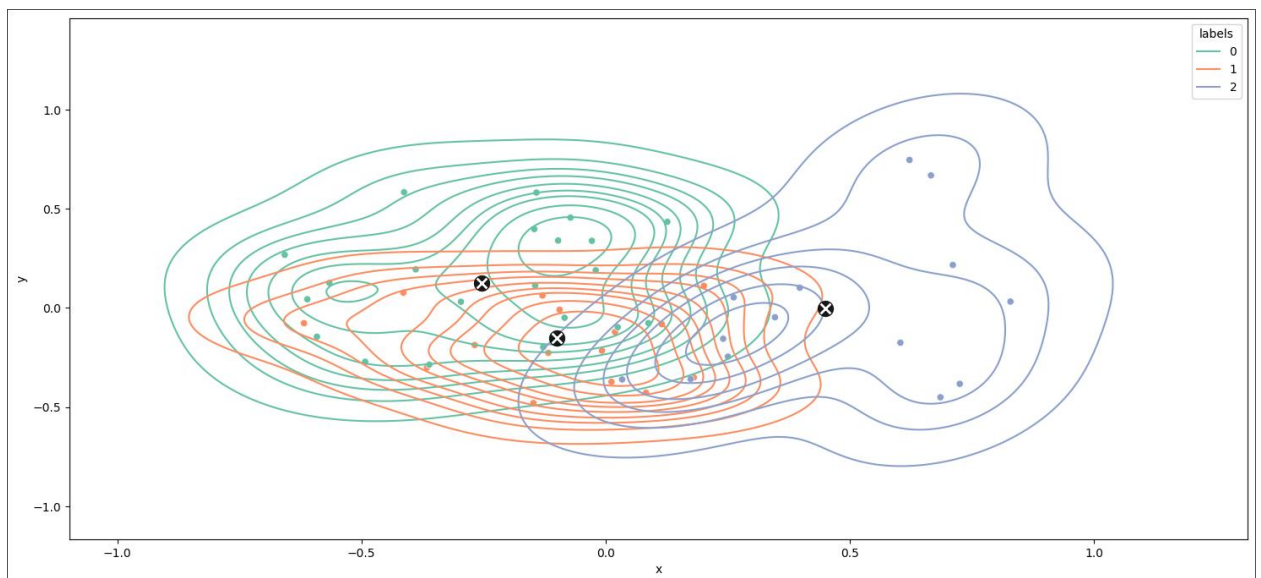


Рисунок 3.5 – Распределение ИТ-проектов между кластерами

Как видно, кластеры довольно сильно пересекаются между собой, что является следствием и демонстрацией «нечеткости», т.е. случаев, когда один и тот-же проект не может быть однозначно отнесен к только одному кластеру. В таких случаях, нельзя однозначно интерпретировать результаты оценки модели, ввиду того, что какие-то из параметров проекта находятся на так называемых «пограничных» значениях, и с определенной вероятностью, может отнесен как к двум кластерам одновременно, так и сразу к трем, что естественно не может быть объективной и однозначной оценкой. Подобные случаи требуют вмешательства со стороны компетентных людей, которые смогут дать более объективную оценку.

Нечеткость результирующей оценки является особенностью работы метода нечеткой кластеризации, что тем не менее позволят провести оценку в условиях, когда другим моделям затруднительно дать хоть какой-то результат, а в особых случаях и вовсе невозможно.

3.2 Тестирование предлагаемой системы

Разработав ML-модель и осуществив ее обучение, воспользовавшись набором данных, который был сформирован и подготовлен экспертной группой, следует провести тестирование, используя «реальный» набор данных, который отличен от использованного во время обучения.

Тестирование моделей сопряжено с определением ключевых показателей и метрик, на основании которых и будет оцениваться модель, ее точность и эффективность.

Параметры ИТ-проектов заданы таким образом, что представляют из себя значения от 0 до 1 и обозначают степень влияния параметра на положительность принятия решения относительно того, подходит ли проект для финансирования, или нет. В рамках разработки методики оценки проектов были выделены следующие варианты результатов работы модели:

- первый кластер – проекты, которые финансировать целесообразно;
- второй кластер – проекты, которые требуют доработки;
- третий кластер – проекты, которые финансировать не целесообразно.

Для оценки точности и корректности работы ML-модели были выбраны следующие метрики:

- средний коэффициент силуэта;
- скорректированный индекс RAND.

Средний коэффициент силуэта – это статистический показатель, который используется для оценки эффективности процесса кластеризации. Он представляет собой меру схожести элементов внутри одного кластера и различия между элементами разных кластеров. Коэффициент силуэта рассчитывается для каждого элемента отдельно, а затем усредняется по всем элементам. Значение среднего коэффициента силуэта может варьироваться от 0 до 1, при этом чем ближе значение к 1, тем выше качество кластеризации.

Скорректированный индекс RAND — это статистическая мера, используемая для оценки качества алгоритмов кластеризации. Она основывается на индексе согласованности Рэнда, но учитывает случайные совпадения между предсказанными и истинными кластерами. Индекс принимает значения от -1 до 1. Значение 1 указывает на идеальное соответствие между предсказанными и истинными кластерами, а значение -1 указывает на полное несоответствие. Значение 0 соответствует случаю, когда кластеризация происходит случайно.

Для реализации оценки работы и точности модели, были использованы функции пакета «sklearn» класса «metrics»: `silhouette_score`, `adjusted_rand_score`.

В рамках проведения оценки метрических показателей модели, были получены следующие результаты тестирования:

- средний коэффициент силуэта – 0.152;
- скорректированный индекс RAND – 0.512.

По итогам тестирования оценки, полученные результаты можно интерпретировать следующим образом:

а) Средний коэффициент силуэта (AS) равен 0.152. Это значение находится в диапазоне от 0 до 1, где 1 является идеальным показателем. Однако, поскольку значение AS близко к 0, это указывает на то, что качество кластеризации может быть улучшено.

б) Скорректированный индекс RAND (ARI) равен 0.512. Этот показатель также находится в диапазоне от -1 до 1, где 1 является идеальным показателем. Значение ARI ближе к 1, что указывает на то, что модель достаточно точно предсказывает кластеры.

Таким образом, модель работает довольно удовлетворительно, что означает возможность дальнейшего внедрения прототипа для опытного использования на реальных данных. Но, с учетом довольно низкого показателя среднего коэффициента силуэта говорит о том, что кластеры весьма кучно расположены, что подтверждается при визуальном отображении результатов распределения кластеров в виде пересекающихся областей. Это означает, что работу модели можно улучшить.

Негативно повлиять на метрические показатели модели могло повлиять недостаточная подготовка данных, или же неоптимальные гиперпараметры.

Кроме метрических показателей, работу модели можно оценить визуально, отобразив в виде графического представления распределения ИТ-проектов на плоскости, которое представлено на рисунке 3.6.

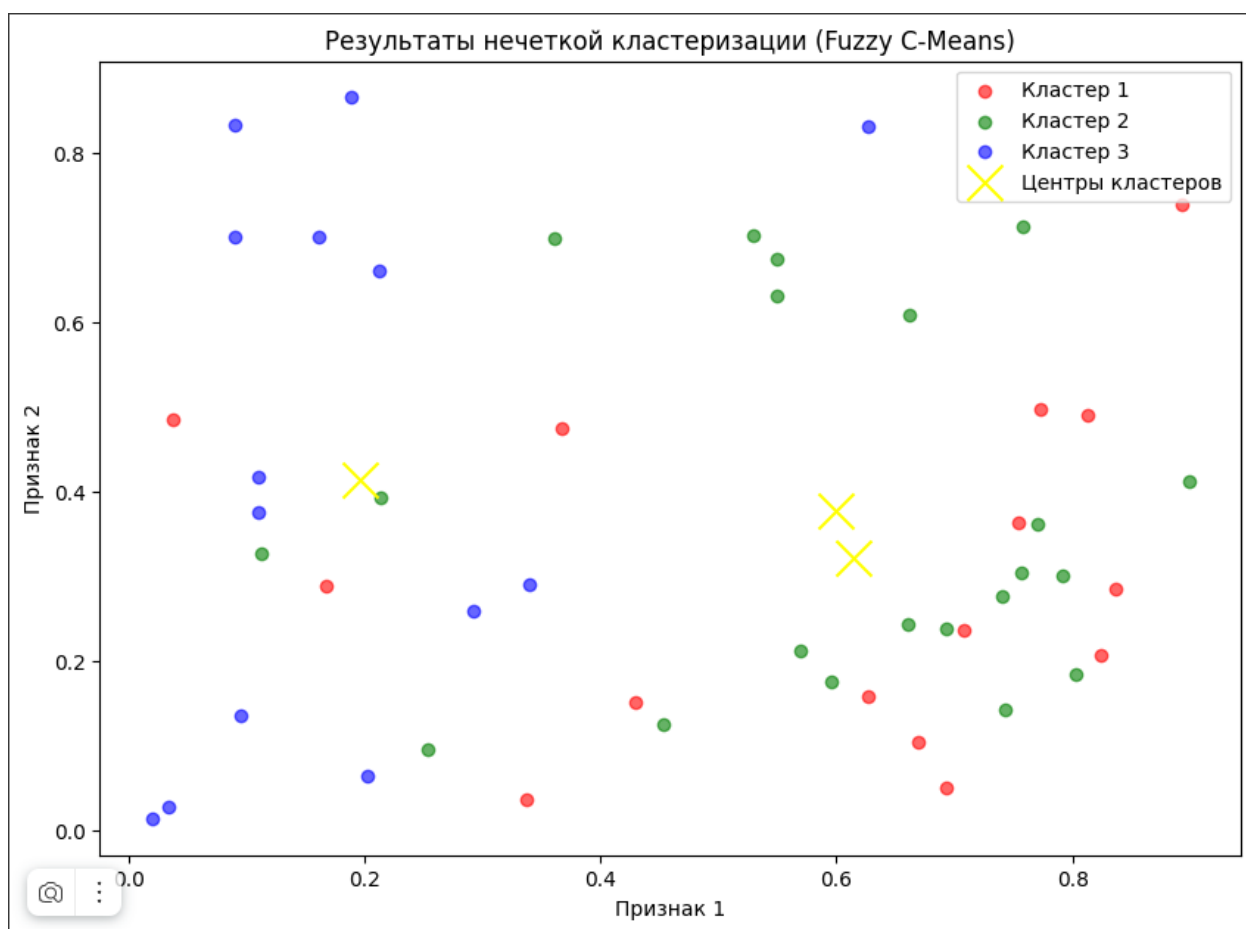


Рисунок 3.6 – Распределение ИТ-проектов, представленное на графике

Как видно на рисунке 3.6, кучность распределения проектов довольно близкая, о чем и говорил показатель среднего коэффициента силуэта, тем не менее, предсказанные моделью значения оказались весьма точными.

3.3 Экономическая оценка перспектив внедрения прототипа системы

Основные перспективы внедрения прототипа системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе заключаются в том, чтобы упростить для региона оценку ИТ-проектов, которые помимо прочих оценочных параметров, оцениваются по степени влияния на критически важную инфраструктуру региона. Данный вопрос принципиален и лежит в основе реализации данной магистерской работы, что обусловлено нестабильной и недружелюбной внешнеполитической

обстановкой. В этих условиях, когда по решению внешних, недружественных стран, которые преследуют собственные интересы, может быть ограничен доступ к тем или иным программным средствам, могут быть ограничены функциональные возможности аппаратных средств, что так или иначе влияет на жизнь населения, экономическое развитие региона, но в особенности может влиять на обороноспособность страны, что недопустимо. Именно поэтому, параметр влияния ИТ-проекта на критически важную инфраструктуру является одним из ключевых.

Кроме этого, основными ключевыми аспектами, позволяющими положительно оценить перспективы внедрения информационной системы, являются:

- снижение рисков – использование модели нечеткой кластеризации позволяет более точно определить инвестиционную привлекательность проектов и снизить риски, связанные с выбором неподходящих проектов для финансирования;

- повышение эффективности – информационная система обеспечивает быстрый и точный анализ данных, что позволяет принимать обоснованные решения при отборе заявок на финансирование;

- создание единого информационного пространства – система обеспечивает эффективное распределение управленческой информации и координацию сбора и хранения данных об инновационных проектах;

- улучшение качества и оперативности принятия решения – система предоставляет возможность анализировать и сравнивать различные проекты, оценивать их экономическую эффективность и группировать по кластерам на основе близости параметрических показателей;

- создание обратной связи – система накапливает базу данных отклоненных и успешно проинвестированных проектов, что позволяет использовать эту информацию для принятия решений при управлении портфелями проектов.

Внедрение прототипа информационной системы анализа данных поможет региональным властям эффективнее отбирать заявки на финансирование цифровых проектов. Это повысит качество принимаемых решений и улучшит управление проектами. Система будет собирать и анализировать данные о проектах, что позволит лучше планировать и контролировать их выполнение.

Для проведения экономической оценки перспектив внедрения прототипа системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе необходимо учитывать следующие аспекты:

а) Стоимость создания и внедрения прототипа системы. Это включает в себя затраты на разработку программного обеспечения, приобретение необходимого оборудования, оплату труда специалистов, участвующих в проекте, и другие сопутствующие расходы.

б) Ожидаемый доход от использования системы. Это может быть связано с увеличением количества одобренных заявок на финансирование, улучшением качества отбираемых проектов, сокращением времени на рассмотрение заявок и другими положительными эффектами.

в) Срок окупаемости проекта. Этот показатель позволяет оценить, через какое время после внедрения система начнет приносить прибыль.

г) Риски и неопределенности. Необходимо учитывать возможность возникновения непредвиденных ситуаций, которые могут повлиять на эффективность работы системы.

Для проведения экономической оценки можно использовать метод чистой приведенной стоимости (NPV), внутренней нормы доходности (IRR) или дисконтированного периода окупаемости (DPP). Каждый из этих методов имеет свои преимущества и недостатки, поэтому выбор зависит от конкретных условий проекта.

Для разработки модели машинного обучения были привлечены:

– младший инженер машинного обучения, с заработной платой 78000 рублей в месяц;

– младший специалист по исследованию данных, с заработной платой 70000 рублей в месяц.

Разработкой модели занимался младший инженер машинного обучения, начиная от написания программного кода и заканчивая обучением и тестированием модели.

Сбором и подготовкой данных занималась экспертная группа в лице младшего специалиста по исследованию данных, который собрал и в дальнейшем обработал набор данных, который в дальнейшем был использован при обучении и тестировании модели.

Разработка модели заняла два месяца работы, что равняется двум месячным окладам разработчика, или 156000 рублей.

Сбор и подготовка данных осуществлялся на протяжении месяца, таким образом стоимость подготовки набора данных равняется месячному окладу младшего специалиста по исследованию данных, или 70000 рублей.

Оценка потенциальной прибыли в случае внедрения и применения данной системы – весьма трудоемкий и сложный процесс, что исходит из специфики направленности работы. Довольно сложно оценить потенциальную прибыль региона, по отбору цифровых проектов, включая импортозамещение и поддержание обороноспособности страны, что выходит за рамки данной выпускной квалификационной работы, поэтому введем допущение и оценим стоимость разработки и внедрения системы, приняв допущение, что на внедрение потребуется месяц работы младшего инженера машинного обучения. Таким образом, стоимость внедрения составит 78000 рублей, или месячный оклад младшего инженера.

Подводя итог, получим стоимость разработки и внедрения, равную 304000 рублей. В общей сложности, на проведение данных работ потребуется 4 месяца, принимая во внимание, что работа ведется последовательно, а не параллельно.

Выводы

В рамках главы, был разработан прототип системы анализа данных для отбора заявок на финансирование цифровых проектов в регионе, с использованием языка программирования Python и следующих пакетов:

- pandas – чтение и работа с табличными данными;
- numpy – работа с векторами;
- fcmeans – реализация нечеткой кластеризации;
- PCA и KernelPCA из sklearn.decomposition – понижение размерности данных;
- seaborn и matplotlib – визуализация полученных результатов;
- silhouette_score и adjusted_rand_score из sklearn.metrics – метрики оценивания модели.

В результате, была получена модель нечеткой кластеризации, для обучения и тестирования которой использовался собранный и подготовленный экспертной группой набор данных. Также, для тестирования и проверки эффективности работы были использованы следующие метрики:

- средний коэффициент силуэта, результат 0.152;
- скорректированный индекс RAND, результат 0.512.

Таким образом, работа модели оценивается удовлетворительно, с учетом возможности дальнейшего совершенствования и развития.

Разработка и внедрение прототипа модели были оценены в 304000 рублей, что включает в себя четыре месяца работы двух специалистов, младшего инженера машинного обучения и младшего специалиста по исследованию данных.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были выполнены следующие задачи:

- изучены основы проектирования систем анализа данных, в рамках чего были проведены анализы существующих методов и предметной области, была описана специфика организации проекта внедрения информационной системы, что позволило обеспечить основу для дальнейшей разработки и построения функциональной системы;

- проведена формализации зависимостей для модели машинного обучения, подготовлена методика оценки инновационных ИТ-проектов, описан процесс сбора и подготовки набора данных, что позволило выбрать метод и алгоритм, для реализации системы отбора заявок на финансирование цифровых проектов;

- определены и построены процессные модели этапов разработки, что позволило определить векторы дальнейшей работы, для реализации информационной системы;

- реализована модель машинного обучения на основе нечеткой кластеризации, при обучении которой применялись данные, собранные и подготовленные экспертной группой;

- проведено тестирование разработанной модели машинного обучения с применением метрик, что позволило оценить точность и эффективность модели, а также выделить потенциальные возможности по улучшению и дальнейшему совершенствованию.

На основании вычисления метрик эффективности модели были сделаны следующие выводы:

- точность модели была оценена довольно высоко, с применением скорректированного индекса RAND, значение которого составило 0.512, что указывает на то, что модель достаточно точно предсказывает кластеры, но имеет место совершенствование;

– кластеры накладываются друг на друга, что может говорить о том, что определение ИТ-проектов может быть затруднено и иметь недостаточно точный результат, из-за вероятности отнесения проекта к отличному от реального кластера, оценка производилась с применением среднего коэффициента силуэта (AS), показатель которого равен 0.152.

Стоимость внедрения прототипа проекта, с учетом стоимости разработки, была оценена в 304000 рублей за 4 месяца работы двух специалистов.

Одним из ключевых моментов выпускной квалификационной работы, следует отметить потенциальное развитие региона и выполнение национального проекта [1]. Кроме того, развитие в сфере импортозамещения, на чем отдельно сделан акцент при подготовке данных, в рамках исполнения указа президента РФ [13]. В совокупности, это позволяет обеспечить дальнейшее экономическое развитие региона и ограждение обороноспособности страны от внешнеполитического влияния недружественных стран [10], за счет замещения иностранного программного обеспечения отечественными разработками, в соответствующих сферах и областях применения.

Таким образом, поставленные задачи были успешно решены, а цель выпускной квалификационной работы достигнута.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Цифровая экономика – Национальный проект «Цифровая экономика» // Национальные проекты России. – URL: <https://национальныепроекты.рф/projects/tsifrovaya-ekonomika> (дата обращения 01.03.2024). – Текст: электронный.
2. Колесов К.И. Финансирование инновационных проектов // ЦИТР. – URL: https://citr.ntnu.ru/upload/FCfinder/files/izdaniya/finans_innov_proect.pdf (дата обращения 01.03.2024). – Текст: электронный.
3. Экономика, финансирование и управление инновационной деятельностью [Текст]: учеб. пособие / А.А. Голубев, А.И. Александрова, М.В. Скрипниченко – СПб.: СПбГУ ИТМО, 2016. – 143 с.
4. Траченко М.Б., Кожанова А.В. Специфика структуры финансирования стартапов на разных стадиях их жизненного цикла [Текст] / М.Б. Траченко, А.В. Кожанова // Финансовый журнал. – 2019. – № 5. – С. 90–103.
5. Сулейманова Н.Р., Демьянова О.В. Региональные факторы реализации проектов по цифровой трансформации. // Научная электронная библиотека «КиберЛенинка». – URL: <https://cyberleninka.ru/article/n/regionalnye-factory-realizatsii-proektov-po-tsifrovoy-transformatsii/viewer> (дата обращения 03.03.2024). – Текст: электронный.
6. Агауров С.Ю., Зыкова Н.В. Особенности и ключевые факторы, оказывающие влияние на реализацию крупных проектов [Текст] / С.Ю. Агауров, Н.В. Зыкова // Экономика и управление. – 2023. – Т. 29. – № 4. – С. 407–413.
7. Зозуля А.В., Зозуля П.В., Еремина Т.Н. Современные проблемы реализации приоритетных национальных проектов // Вестник Евразийской науки. – URL: <https://esj.today/PDF/21ECVN119.pdf> (дата обращения 22.05.2024). – Текст: электронный.

8. Инновационно-инвестиционный анализ и оценка проектов [Текст]: учеб. пособие, 2-е изд. / И.А. Бурса, О.В. Тахумова – Краснодар: ООО «Принт-Терра», 2021. – 113 с.

9. И все-таки про оценку проектов // vc.ru – бизнес, технологии, идеи, модели роста, стартапы. – URL: <https://vc.ru/dev/197233-i-vse-taki-pro-ocenku-proektov> (дата обращения 04.03.2024). – Текст: электронный.

10. Обухова А.С., Щербаченко Е.Р., Гусельникова Л.Н. Оценка эффективности инновационных проектов // Вестник Северо-Кавказского федерального университета. – URL: <https://dspace.ncfu.ru/bitstream/20.500.12258/22349/1/B%2032-39%20%282%29.pdf> (дата обращения 20.05.2024). – Текст: электронный.

11. Спешилова Н.В., Сгибнева А.Ю. Применение математических методов к оценке эффективности инвестирования агропромышленного комплекса Оренбургской области // Интернет-журнал «Науковедение». – URL: <http://naukovedenie.ru/PDF/54EVN615.pdf> (дата обращения 07.03.2024). – Текст: электронный.

12. Блинова Е.А. Алгоритм и математические модели оценки инвестиционного проекта предприятия по оказанию услуг // Российский экономический Интернет-журнал – EREJ. – URL: <https://www.e-rej.ru/Articles/2018/Blinova.pdf> (дата обращения 23.05.2024). – Текст: электронный.

13. Математические методы и модели управления проектами [Текст]: учеб. пособие / И.В. Буркова, Я.Д. Гельруд, О.В. Логиновский, А.Л. Шестаков – Челябинск: Издательский центр ЮурГУ, 2018 – 193 с.

14. Крыжановский О.В., Чулюков В.А. Сравнение математических моделей оценки эффективности проектных решений // Современные проблемы науки и образования (сетевое издание). – URL: <https://science-education.ru/ru/article/view?id=19840> (дата обращения 20.05.2024). – Текст: электронный.

15. Яшин С.Н., Коробова Ю.С., Захарова Ю.В. Проблемы финансирования и результативности инновационной деятельности // Журналы Самарского университета. – URL: <https://journals.ssau.ru/eco/article/download/10181/8953> (дата обращения 10.03.2024). – Текст: электронный.
16. Проблема финансирования инновационной деятельности в России на современном этапе [Текст] / Е.Е. Складорова // Научно-методический электронный журнал «Концепт». – 2017. – Т. 2. – С. 296–298.
17. Казанская А.А., Мишура Л.Г. Использование машинного обучения в инвестиционной деятельности // Научный журнал НИУ ИТМО. – URL: <https://economics.ihbt.ifmo.ru/file/article/19683.pdf> (дата обращения 11.03.2024). – Текст: электронный.
18. Machine Learning (in Finance) // Corporate Finance Institute – URL: <https://corporatefinanceinstitute.com/resources/data-science/machine-learning-in-finance/> (дата обращения 23.05.2024). – Текст: электронный.
19. Искусственный интеллект в финтехе и банкинге // ГАРАНТ – Законодательство (кодексы, законы, указы, постановления) РФ, аналитика, комментарии, практика. – URL: <https://www.garant.ru/article/1649470/> (дата обращения 24.05.2024). – Текст: электронный.
20. 10 Best Use Cases of Machine Learning in Finance // Fayrix | Offshore Custom Software Development Services. – URL: <https://fayrix.com/blog/machine-learning-in-finance> (дата обращения 24.05.2024). – Текст: электронный.
21. ML Architecture // Evaluation for LLM-Based APPS | Deepchecks. – URL: <https://deepchecks.com/glossary/ml-architecture/> (дата обращения 25.05.2024). – Текст: электронный.
22. Machine Learning Architecture It Is, Component & Types // Git for Data – lakeFS. – URL: <https://lakefs.io/blog/machine-learning-architecture/> (дата обращения 20.05.2024). – Текст: электронный.

23. Архитектура платформы машинного обучения в продакшене // vc.ru – бизнес, технологии, идеи, модели роста, стартапы. – URL: <https://vc.ru/ml/537806-arhitektura-platformy-mashinnogo-obucheniya-v-prodakshene> (дата обращения 15.03.2024). – Текст: электронный.

24. Роль математики в машинном обучении // Публикации / Моя лента / Хабр. – URL: <https://habr.com/ru/companies/otus/articles/525636/> (дата обращения 23.04.2024). – Текст: электронный.

25. What is the significance of mathematics in Machine Learning? // KnowledgeHut: Professional Bootcamps and Certification Courses Provider for your Future. – URL: <https://www.knowledgehut.com/blog/data-science/importance-of-maths-in-machine-learning> (дата обращения 17.04.2024). – Текст: электронный.

26. Machine Learning Mathematics // A computer science portal for geeks. – URL: <https://www.geeksforgeeks.org/machine-learning-mathematics/> (дата обращения 18.05.2024). – Текст: электронный.

27. Об утверждении перечня иностранных государств и территорий, совершающих недружественные действия в отношении Российской Федерации, российских юридических и физических лиц [Текст]: Распоряжение Правительства РФ от 05.03.2022 г. № 430-р (ред. от 29.10.2022) // Официальный интернет-портал правовой информации. – 2022. – 7 марта.

28. Кластеризация данных при помощи нечетких отношений в Data Mining // Министерство Образования и Науки Российской Федерации. – URL: https://ami.nstu.ru/~vms/lecture/data_mining/fuzzy.htm (дата обращения 23.04.2024). – Текст: электронный.

29. Ранжирование объектов на основе нечеткой кластеризации / Н.Д. Баданина, А.А. Зинченко, В.А. Судаков // Препринты ИПМ им. М.В.Келдыша, 2022. – № 68. – 12 с.

30. Fuzzy Clustrering in R // GeeksforGeeks | A computer science portal for geeks. – URL: <https://www.geeksforgeeks.org/fuzzy-clustering-in-r/> (дата обращения 24.04.2024). – Текст: электронный.

31. Акишев А.А., Близоруков М.Г. Применение методов нечеткого кластерного анализа к задаче сегментации потребителей рынка изоляционных материалов // CEUR Workshop Proceedings (free, open-access publishing, computer science/information systems). – URL: <https://ceur-ws.org/Vol-1825/p25.pdf> (дата обращения 03.05.2024). – Текст: электронный.

32. Нечеткая логика – математические основы // Loginom. – URL: <https://loginom.ru/blog/fuzzy-logic> (дата обращения 02.05.2024). – Текст: электронный.

33. О мерах по обеспечению технологической независимости и безопасности критической информационной инфраструктуры Российской Федерации [Текст]: Указ Президента Российской Федерации от 30.03.2022 г. № 166 // Официальный интернет-портал правовой информации. – 2022. – 30 марта.

34. Теория и практика машинного обучения [Текст]: учеб. пособие / В.В. Воронина, А.В. Михеев, Н.Г. Ярушкина, К.В. Святков – Ульяновск: УлГГУ, 2017. – 290 с.

35. Машинное обучение на практике. Опыт применения облачных технологий и машин-лернинга в крупном проекте // Хакер – Безопасность, разработка, DevOps. – URL: <https://xakep.ru/2018/08/01/rts-tender/> (дата обращения 25.04.2024). – Текст: электронный.

36. Применение машинного обучения для решения задач в IT-отрасли: практические примеры и перспективы развития // SecurityLab.ru. – URL: <https://www.securitylab.ru/analytics/538199.php> (дата обращения 05.05.2024). – Текст: электронный.

37. ML | Fuzzy Clustering // GeeksforGeeks | A computer science portal for geeks. – URL: <https://www.geeksforgeeks.org/ml-fuzzy-clustering/> (дата обращения 02.05.2024). – Текст: электронный.

38. Проблемы имплементации искусственного интеллекта в сфере образования / Л.Н. Духанина, А.А. Максименко // Перспективы науки и образования, 2020. – № 4 (46). – С. 23–35.

39. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden // BMC Health Services Research. – URL: <https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08215-8> (дата обращения 10.05.2024). – Текст: электронный.

40. ML не в радость: что может провалить проект по внедрению machine learning // КОРУС Консалтинг – аналитические решения, продвинутая аналитика и управление данными. – URL: <https://data.korusconsulting.ru/press-center/blog/ml-ne-v-radost-chno-mozhet-provalit-proekt-po-vnedreniyu-machine-learning/> (дата обращения 27.04.2024). – Текст: электронный.