Санкт–Петербургский государственный университет

ГОЛОВИНА Светлана Владимировна

Выпускная квалификационная работа

Применение методов машинного обучения в задаче определения уровня рисков во время беременности

Уровень образования: бакалавриат
Направление 02.03.02 «Фундаментальные информатика и информационные технологии»
Основная образовательная программа СВ.5003.2017 «Программирование и информационные технологии»
Профиль «Автоматизация научных исследований»

Научный руководитель:
доцент, кафедра компьютерного моделирования и многопроцессорных систем, к.ф. - м.н.
Корхов Владимир Владиславович

Рецензент:
генеральный директор общества с ограниченной ответственностью «Виста»
Савватеев Александр Юрьевич

Санкт-Петербург
2021 г.

Содержание

Введение

Рост достижений в сфере информационных технологий положительно сказывается на развитии практически всех областей знаний. Разработка и внедрение информационных систем в нашу жизни являются на сегодняшний день одними из самых актуальных задач.

В период пандемии на медицинские учреждения и их работников сильно возрасла нагрузка, поэтому использование автоматизированных систем, упрощающих работу персонала, приобрело особое значение. Это дало толчок к развитию направления дистанционных медицинских консультаций, позволяющих поддерживать связь пациентов с медработниками и сократить количество личных обращений в больницы. Для оптимизации времени ожидания от такой системы полезна программа, которая могла бы проанализировать ответы пациента на простые вопросы о самочувствии и автоматических вывести по ним предварительную оценку состояния человека.

Такая задача осложняется тем, что в ней невозможно учесть все имеющиеся условия, влияющие на ответ, — можно лишь выделить примерный набор наиболее важных признаков. Полученный результат при этом будет носить только приблизительный характер, а алгоритм его нахождения не может быть выписан точно и последовательно [1].

В приложении «ТАДАМ» компании «Виста» уже организуются индивидуальные комнаты консультаций с врачом с помощью чат-ботов, позволяющие получить рекомендации, не посещая больницу. Удаленные консультации со специальными возможностями для беременных женщин являются следующим шагом развития продукта.

Постановка задачи

Целью данной работы является исследование методов машинного обучения на данных из электронных медицинских карт для применения результатов в системе удаленных консультаций беременных женщин. Такая система должна обрабатывать доступную информацию по каждому пациенту и выдавать результат — уровень опасности состояния. В зависимости от полученного показателя будут рассчитываться электронная очередь на консультации и экстренные перенаправления к врачам. Таким образом планируется предотвратить критические задержки в оказании помощи.

Для реализации планируемой задачи необходимо:

1. Проанализировать способ организации записей в электронных медицинских картах из предоставленной базы данных.

2. Автоматизировать сбор необходимой информации из электронных медицинских карт.

3. Определить методы обработки полученных данных и сформировать представления пациентов.

4. Создать прототип системы, классифицирующей состояние пациента, основываясь только на описательных данных.

Обзор литературы

В обзоре рассмотрены четыре исследования, по-разному решающих поставленную или схожую задачу. Все они выбраны, так как раскрывают достоинства и недостатки наиболее распространенных подходов. Анализ представленных работ помог обозначить перспективное решение рассматриваемой задачи.

Работа [2][3] является примером легитимной дистанционной медицинской врачебной помощи и акушерского мониторинга, разработанного в соответствии с техническим заданием Министерства здравоохранения Свердловской области. Основой данного метода является унифицированная анкета оценки группы риска, которая автоматически рассчитывает потенциальный общий перинатальный и глобальные риски пациентки. В статье [3] не описан конкретный метод оценки, но указано, что используются электронный «Бенчмаркинг» – эталонное тестирование состояния здоровья согласно существующим протоколам лечения, и технология «Глобальные риски» — «Ноу-хау» специалистов службы родовспоможения Свердловской области. То есть в данной работе задача медицинской диагностики решена в соответствии с существующими клиническими рекомендациями, требующими профессиональной медицинской подготовки.

В диссертационной работе [4] рассмотрена бионическая модель для решения задачи мониторинга состояния здоровья беременных женщин-плода-детей. Разработана информационная структура, реализующая интегрированное представление знаний об объекте исследования и его функционировании. Бионическая модель структурно представляет собой комплекс взаимосвязанных средств обработки внутренней и внешней информации, основанный на генетическом и нейросетевых алгоритмах. Метод вычисления обобщенного показателя биосистемы Мать-Плод формирует единый интегральный показатель для слежения за динамикой изменения состояния и адаптационными характеристиками биосистемы. Искусственные нейронные сети позволяют получить прогноз интегральной оценки состояния объекта исследования в зависимости от выбранных управляющих воздействий. Генетический алгоритм осуществляет выбор последовательности управля-

ющих воздействий, которые снижают возможность перехода в неблагоприятное состояния.

Одним из решений телемедицины является “SF Medic” [5][6], созданный на основе Amazon Web Services и предназначенный для замены несерьезных посещений больниц. Он обеспечивает поддержку принятия клинических решений в режиме реального времени через предупреждения, такие как потенциальные противопоказания к лекарствам, с помощью Amazon Comprehend Medical, Amazon Transcribe Medical и Amazon Translate. Наибольший интерес представляет Amazon Comprehend Medical, используемый для извлечения медицинских терминов из текстовых сообщений. Он работает на предварительно обученной модели обработки естественного языка и позволяет изучать клинические документы, чтобы получить информацию об их содержании для анализа неструктурированного клинического текста посредством обнаружения сущностей. Сущность — это текстовая ссылка на медицинскую информацию, такую как состояние здоровья или лекарства. Создатели не уточняют, какой именно метод при этом используют, но наличие такой разработки уже позволяет делать выводы, что у методов обработки естественного языка есть потенциал в сфере телемедицины.

Целью работы [7] является преобразование разнородных клинических данных из электронных медицинских карт в клинически значимые конструктивные признаки с помощью метода, который частично опирается на временные отношения между данными. На основе записей о 265336 пациентах с 555609 уникальными клиническими событиями строятся векторные представления медицинских терминов, поддерживающих семантически нагруженные линейные операции, то есть сохраняющих взаимосвязи терминов. Алгоритм Skip-gram, лежащий в основе рассматриваемой работы, фиксирует отношения между словами и каждому медицинскому термину предсказывает его контекст на основе частоты совместной встречаемости в электронных медицинских картах. По таким векторам агрегируется векторное представление пациента и для сравнения обучаются 4 модели предсказания вероятности наступления сердечной недостаточности. Особый интерес к этому исследованию возникает из-за того, что в нем не использовались никакие экспертные знания и его потенциально можно

применять не только для предсказания заболеваний сердца, но и в других областях медицины.

В статье также упоминается, что большинство существующих подходов не использует непосредственно данные электронных медицинских карт, а реализуют поиск сущностей и их анализ. Авторы предложили новый способ представления гетерогенных медицинских концепций на основе паттернов совместного возникновения в продольных электронных медицинских записях.

Анализ литературы показал актуальность исследований в области применения методов машинного обучения к мониторингу беременных женщин и телемедицине. Первые два рассмотренных метода показывают очень хорошие результаты, зарекомендовав себя и найдя применения в больница. В рамках данной работы их недостатками являются требования к наличию экспертных знаний и большого объема данных, которые можно получиться только во время посещения медицинского учреждения (анализы крови, ЭКГ и т.д.). Третье и четвертое исследования показывают перспективы применения методов обработки естественного языка для задач медицинского профиля. Их достоинством для исследования является то, что в отличии от двух других, анализ данных не требует специальных медицинских знаний и основывается только на обработке историй болезней из электронных медицинских карт.

В настоящем исследовании учтены результаты анализируемых работ и сферы их применений. Подход к представлению данных, описанный в статье [7], лег в основу исследования.

Глава 1. Анализ и извлечение данных

1.1 Описание структуры БД

Компания «Виста» занимается оптимизацией деятельности медицинских организаций с помощью электронных баз данных. Одна из них рассматривается в данной работе. Все записы были обезличены и получены за период 2011-2020 годов.

В предоставленной базе данных связь клиентов и информации об их историях болезней осуществляется через 5 таблиц:

1. «Client» хранит информацию об id клиентов, половой принадлежности и дате рождения;

2. «Event» связывает id клиентов и id каждого их посещения;

3. «Action» соединяет id каждого посещения с id каждой предоставляемой услуги (процедуры, приема и т.д.);

4. «ActionProperty» связывает id услуги со всей полученной на ней информацией;

5. «ActionPropertyString» имеет тот же id, что и «ActionProperty», но содержит только описательную информацию о состоянии клиента. Именно она представляет наибольший интерес для анализа.

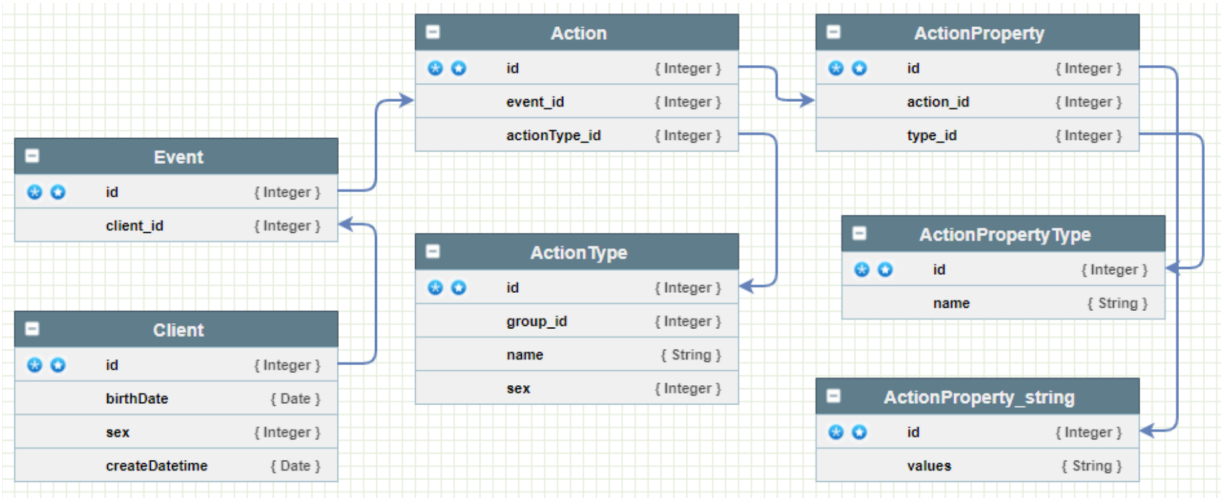

Рис. 1: Схема используемой части предоставленной базы данных.

Также для лучшего понимания принципа хранения данных использовались таблицы «ActionPropertyType» и «ActionType», содержащие назва-

ния оказываемых услуг и группы, к которым они относятся. Именно эти данные помогли отобрать необходимую для работы записи.

Структура используемой в работе части базы данных представлена на рисунке 1.

1.2 Определение критериев отбора записей

Чтобы среди всех клиентов выбрать только тех, кто когда-либо состоял на учете по беременности, был проведен анализ данных о типах приемов. Наиболее полезной оказалась таблица «ActionType», которая частично описывает действия, оказываемые пациентам. В частности, она содержит информацию о специфичности половой принадлежности для каждой услуги:

- 0 — для нейтральной;
- 1 — исключительно для мужчин;
- 2 — исключительно для женщин.

С помощью этих данных получилось выделить 5 групп действий, оказываемых только женщинам. Предполагалось, что эти записи связаны с беременностью.

При их дальнейшем изучении [8] обнаружилось, что:

- группа «111» не всегда относится к записям о беременности, в общем случае это данные о стационарном лечении. Интересующим нас оказался только один тип действий из этой группы, относящийся к поступлению в родовое отделение («62347»). В дальнейшем записи об этой группе не учитываются, так как цель данной работы в дистанционном отслеживании состояния во время беременности до поступления в стационар. Группа «152» является парной с «111» — осмотр перед поступлением в медицинское учреждение. Также не учитывается.
- группу «157» перестали использовать, поэтому она не была включена в дальнейшее исследование.

- для остальных двух групп («692», «695») предположение о записях, относящихся только к беременности, подтвердилось. Они легли в основу отбора клиентов.

Из описанных выше двух групп были собраны id анамнезов, так как они имели схожую структуру и наиболее полный набор данных для анализа. Именно номер действия стал ключевым критерием отбора записей.

1.3 Сбор данных

Большую роль в обработке электронных медицинских карт играют текстовые и числовые медицинские данные, находящиеся в выписках пациентов. Большинство из них имеют описательный характер и выражаются с помощью медицинских терминов и сокращений, имеющих различных вариации. В связи с этим существует потребность в их обработке и анализе.

По найденному критерию номеров действий был произведен первичный отбор id клиентов, когда-либо стоявших на учете по беременности, с помощью SQL запроса [9]. Получилось 39872 пациента.

Так как база составлялась вручную работниками медицинских учреждений, была произведена очистка от записей, вызывающих сомнения в достоверности (неправильная половая принадлежность, отсутствие данных о диагнозе и т.д.). Всего получилось 154348 записей.

Следующим этапом было выделение и группировка доступных данных. Записи о посещениях имеют различную структуру, что усложнило формирование ключевых признаков. После проверки нескольких вариантов были выбраны следующие показатели:

- Возраст — разница даты обращения и рождения;
- Срок беременности в неделях — поиск регулярным выражением в диагнозе;
- Триместр - определяется из срока;
- Жалобы — отдельное поле в записях о приеме;

- Диагноз — отдельное поле в записях о приеме;
- Опасные факторы — объединение полей о наследственности, вредных привычках, данных о муже;
- Описательные данные — объединение полей об осмотре, гинекологических особенностях, течении беременности;
- Данные о предыдущих беременностях и заболеваниях;

Выполняемая задача должна использоваться для составления очереди на электронные консультации, поэтому планируется, что изменяющаяся информация (возраст, срок, жалобы) определяется самим пациентом в разделе удаленных консультаций, а остальные данные собираются с последнего приёма с помощью SQL запроса, так как большинство из них не меняется.

После очистки данных была проведена разметка записей на 3 уровня опасности в соответствии с оценками пренатальных факторов риска [10]. Пример собранных данных приведен на рисунке 2.

	weeks	com	diagnos	risk_st	fir	sec	thr	trim1	trim2	trim3	age1	age2	age3
0	5.0	нет	оаг аллерг пенициллин	0	пенициллин крапивниц не_применя смотрет тон се...	здоров	веден согласн приказ 572н минздравсоцразвит к...	1	0	0	1	0	0
1	9.0	нет	синдр потер плод рубец матк кесар сечен шеечн ...	2	не_применя партнер тон сердц ясн везикулярн не...	ребенок впс умерет порок здоров	веден согласн приказ 572н минздравсоцразвит о...	1	0	0	0	1	0
2	10.0	нет	рубец матк кесаркв сечен гипотиреоз наследстве...	1	не_применя скоп гистероскоп взомт эндометриоз ...	здоров	обследов мгк возрастн уз щитовидн желез димер...	1	0	0	0	1	0
3	10.0	нет	ретрохориальн гематом седловидн матк хроническ...	1	тиосульфат натр потер сознан тахикард не_приме...	инсульт здоров	веден согласн приказ 572н минздравсоцразвит у...	1	0	0	0	1	0
4	6.0	бол вниз живот сохраня маза кровянист выдел пр...	угроз самопроизв выкидыш миом матк интерст лок...	2	барьерн не_тон сердц ясн везикулярн не_увелич ...	инфарт сах диабет срок учет детств не_состоя г...	обследов приказ гормон щитовидн желез ттг гом...	1	0	0	0	1	0

Рис. 2: Пример данных, собранных из электронных медицинских карт.

Глава 2. Представление данных

2.1 Предобработка текстовых признаков

Исходя из того, что в поставленной задаче анализ должен производиться на основе произвольных ответов и записей, имеющих разную структуру, использование методов обработки естественного языка имеет наибольший потенциал.

Работа с выделенными данных начинается с очистки и подготовки корпуса, чтобы уменьшить влияние грамматической составляющей на дальнейший анализ [11]. Предварительная обработка данных состоит из следующих этапов:

1. Очистка от пунктуации, цифр и сведение к одному регистру. Реализовано через регулярные выражения и строковый метод lower().

2. Токенизация - перевод текста в список текстовых единиц (слов). Также были учтены слова с частицей не, они были объединены в один токен для сохранения смысла.

3. Удаление стоп-слов - наиболее часто встречающихся токенов, не несущих смысловой нагрузки (союзы, предлоги и т.д.). Позволяет уменьшить словарь текста и уравновесить значения важных информативных слов. Использовался словарь русского языка, предоставляемый библиотекой nltk, дополненный частыми словами, не относящимися к оценке состояния здоровья (названия больниц, городов и т.д.).

4. Лемматизация и стемминг. Чтобы уменьшить корпус схожих слов, применялись методы определения словоформ: стемминг - перевод слова к нормальной форме или корню, лемматизация - приведение к канонической форме слова. Для русского языка есть уже подготовленные функции стемминга и лемматизации. Для этого использовались модуль SnowballStemmer из библиотека nltk.stem и pymorphy2. В представленных данных очень много опечаток и специальных терминов, поэтому используются оба подхода последовательно, так как применение одного из них не значительно уменьшает количество схожих слов.

2.2 Векторное представление текстовых признаков

Большинство алгоритмов машинного обучения подразумевают представление данных в виде вещественных векторов одинаковой размерности, поэтому очищенные данные необходимо векторизовать [12].

Векторизация - это процесс кодирования текста в виде числовой формы для создания векторов признаков. В данной работе использовался подход вложений (embeddings). Он позволяет создать представления текстов кодированием контекстной информации, основываясь на идеи, что слова, появляющиеся в схожем контексте, будут ближе в векторном пространстве. Таким образом вложения представляют собой отображения слов в соответствующие n-мерные векторы.

В данной работе рассмотрены подходы Word2Vec [13], FastText [15] и BertEmbedings [17].

2.2.1 Word2Vec

Модель Word2Vec использует неглубокую нейронную сеть для построения вложений слов на основе максимизации косинусной близости между векторами слов, которые часто появляются друг с другом, и ее минимизации между словами, которые не появляются рядом. Существует две возможные реализации данного метода: Continuous Bag of Words (CBOW) и Skip-gram, архитектура которых представлена на рисунке 3. На вход подается корпус текста, а на выходе получается набор векторов слов. CBOW строится на предсказании слова при данном контексте, а Skip-gram наоборот — предсказывает контекст при данном слове.

У данного подхода есть возможность использовать предобученные векторные представления или проводить обучение на собственных текстах. Так как электронные медицинские карты содержат много специфических терминов, а объем собранных данных достаточно большой, был выбран подход самостоятельного обучения модели с помощью библиотеки gensim [14]. Полученная модель содержит более точные и полезные связи между словами данной тематики. Визуализация примера вывода близких слов приведена на рисунке 4.

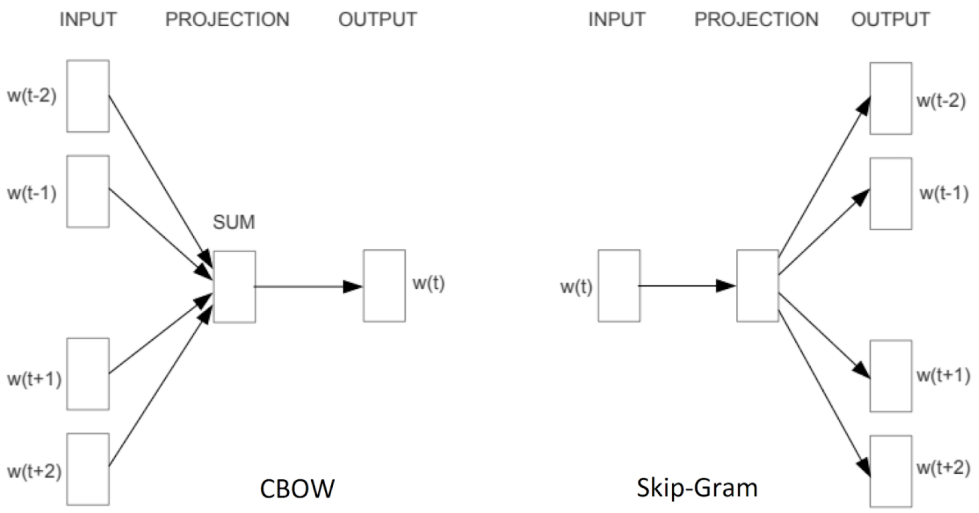

Рис. 3: Архитектура Word2Vec.

Параметры обучения: skip-gram, vector size=300, window=5, negative sampling=7, min count =10.

Особенностью представленного подхода является то, что слова вне обученного словаря не учитываются при построении векторного представления. С одной стороны, это помогает уменьшить влияние незначимых и редких слов. С другой стороны, такое решение игнорирует слова с опечатками, упуская возможно важные слова.

2.2.2 FastText

FastText — неглубокая нейронная сеть с одним скрытым слоем. Её можно считать расширением Word2Vec, так как в его основе лежат такие же идеи и схожая реализации, но добавлена модель символьных n-грамм. Каждое слово представляется композицией нескольких последовательностей символов определённой длины, а вектор слова - суммой всех его n-грамм. Такой подход позволяет обрабатывать и генерировать векторные представления слов, которые модель ранее не встречала, что является преимуществом перед Word2Vec.

FastText также имеет подготовленную модель с предобученными векторными представлениями слов. Чтобы сравнить результаты самостоятельно обученных моделей [16] и готовых векторов, были реализованы оба под-

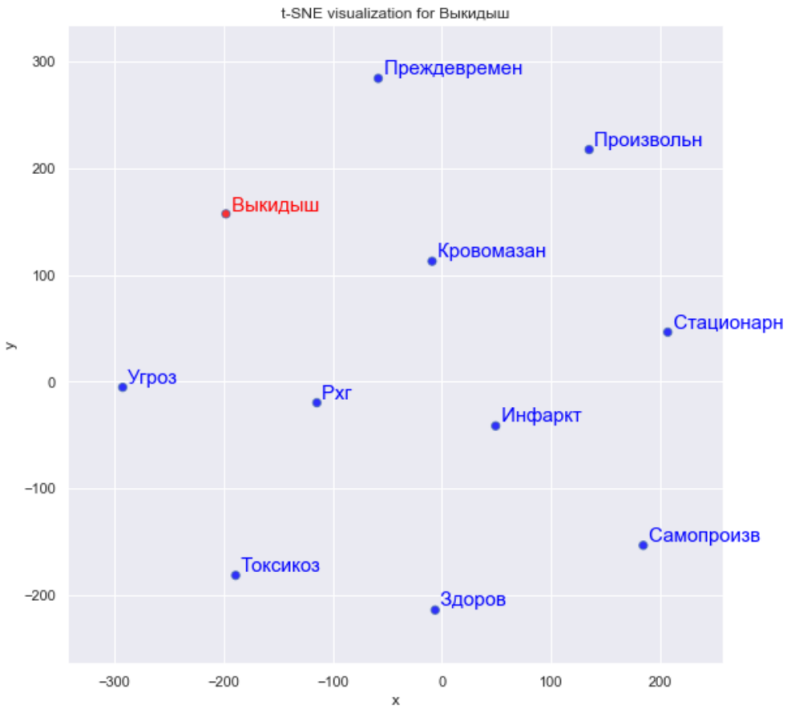

Рис. 4: Пример вывода соседних векторов слов в модели W2V.

хода. Параметры обучения: skip-gram, vector size=300, window=5, min count =10, n gram=2-4, epochs=30. Визуализация примера вывода близких слов для самообученной модели приведена на рисунке 5.

Если сравнить рисунки 4 и 5, представляющие наиболее близкие слова к слову “выкидыш”, то видно главное различие моделей - обработка редких слов. FastText учитывает все слова и определяет их сходство даже с серьезными ошибками в написании.

2.2.3 BertEmbedings

BERT (от Bidirectional Encoder Representations from Transformers) — языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в задачах обработки естественного языка. Технология, основанная на нейросетях, помогает понимать и обрабатывать естественный язык.

В отличие от классических языковых моделей, BERT обучает контекстно-зависимые представления. Он учитывает окружающий контекст предложе-

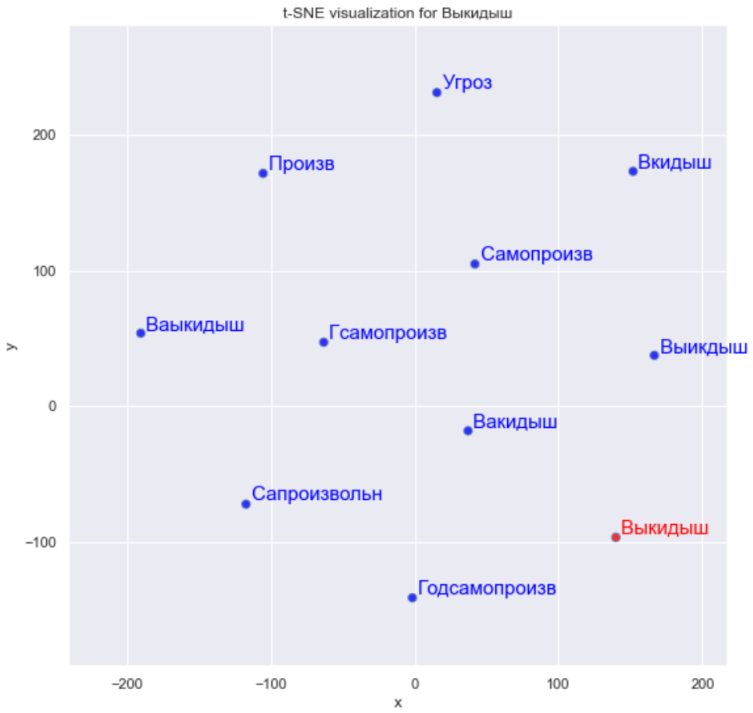

Рис. 5: Пример вывода соседних векторов слов в модели FastText.

ния и генерирует различные векторные представления для разных значений одного и того же слова, что является преимуществом перед Word2Vec и FastText.

В данной работе используется RuBERT (Russian, cased, 12-layer, 768-hidden, 12-heads, 180M parameters), обученный на русской части Wikipedia и новостных данных [18].

2.3 Векторное представление категориальных признаков

Важными признаками, влияющими на оценку состояния, являются возраст и триместр. Они не являются текстовыми, но тоже требуют обработки:

- При анализе возраста важно учитывать только меньше ли он 18 (ранняя) или больше 40 (возрастная);
- Триместр является указателем существенных изменений в состоянии беременной, так как норма описательных данных на разных сроках

сильно отличается;

Получаем, что сами значения возраста и триместра не несут в себе информацию. Они помогают разделять данные по разным группам с соответствующими особенностями, то есть являются категориальными признаками [19]. Чтобы алгоритмы правильно интерпретировали категориальные признаки их необходимо закодировать. В данном случае оба признака имеют всего 3 значения, поэтому использовалось преобразование методом one hot encoding, в котором все элементы в векторе равны 0, кроме одного, определяющего принадлежность к соответствующей категории. Принцип кодирования изображен на рисунке 6.

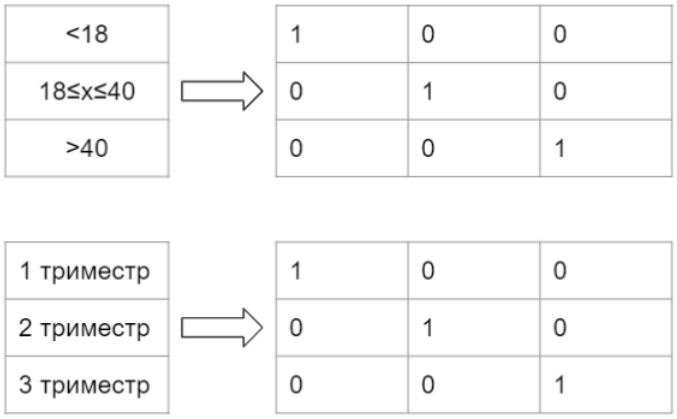

Рис. 6: Кодирование категориальных признаков.

2.4 Векторное представление пациента

После векторизации текстовые признаки представляют из себя матрицы размера (300, n) для Word2Vec и FastText или (768, n) для BertEmbedings, где n - количество слов. Такой формат не применим для дальнейших алгоритмов, поэтому вектора слов суммируются и делятся на их количество - n. Получаем вектор, отражающий среднее значение текстового признака.

Последовательная конкатенация всех описанных выше признаков приводит к созданию вектора пациента, изображенного на рисунке 7.

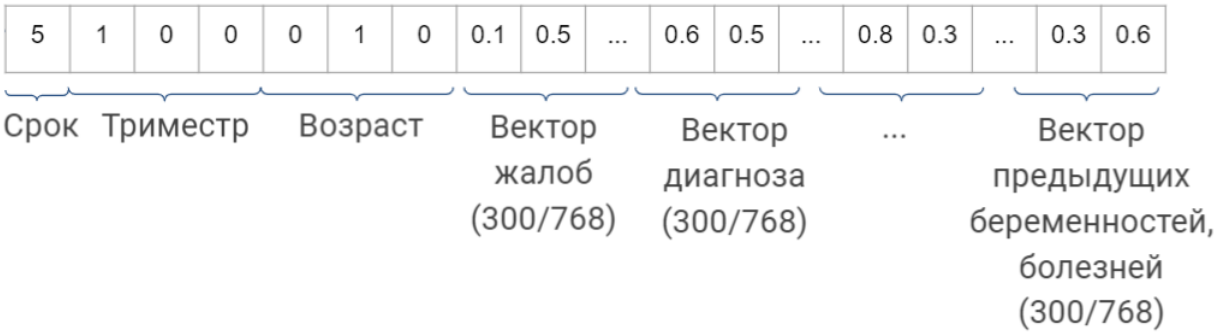

Рис. 7: Кодирование категориальных признаков.

Глава 3. Определение уровня риска

3.1 Методы классификации

Задача определения уровня рисков во время беременности сводится к задаче классификации. X — множество объектов — векторное представление пациента, Y представляет собой конечное множество классов — оценка состояния от 0 до 2, где 0 - минимальные риски, 2 - серьезная угроза. Требуется построить такой алгоритм a: X $\to$ Y, который любому объекту x$\in$X ставит в соответствие метку класса y$\in$Y [20].

В качестве алгоритмов классификации были использованы методы K-ближайших соседей, опорных векторов, случайный лес и логистическая регрессия.

K-ближайших соседей (KNN) - это метрический алгоритм, относящий классифицируемый объект к тому классу, к которому принадлежит наибольшее из k ближайших обучающих объектов. Он сохраняет экземпляры обучающых данных и сравнивает классифицируемый объект со всеми объектами выборки, не строя общую внутреннюю модель [21]. Для работы алгоритма необходимо хранить всю обучающую выборку, что приводит к неэффективному расходу памяти и усложнению решающего правила. Настраиваемый параметр - число соседей k [22].

Метод опорных векторов (SVM) - линейный классификатор, основанный на построении оптимальной разделяющей гиперплоскости для каждого класса. Обучение SVM сводится к задаче квадратичного программиро-

вания, имеющей единственное решение, которое вычисляется достаточно эффективно даже на большин выборках [20]. Для функции принятия решения могут быть указаны различные функции ядра, увеличивая универсальность подхода [23]. Настраиваемый параметр - размер ядра [24][25].

Случайный лес (RF) - ансамблевый алгоритм, строящий деревья решений для случайных подвыборок данных и определяющий класс путем усреднения их вероятностного прогноза [26], основанного на поиске конъюнктивных закономерностей [27]. Случайный лес наследует большинство достоинств деревьев решений и решает часть недостатков: использование подвыборок и объединение деревьев позволяет уменьшить дисперсию оценки леса и снизить переобучение. Настраиваемые параметры: количество деревьев, максимальная глубина, минимальное количество выборок в ведущем узле, минимальное количество выборок в листовом узле [28].

Логистическая регрессия (LR) - это статистический линейный алгоритм, прогнозирующий вероятность принадлежности к классу с помощью логистической кривой [20]. Из этого следует дополнительная возможность получать численные оценки вероятности принадлежности каждому из классов. Настраиваемые параметры: алгоритм обучения и гиперпараметр, обратный коэффициенту регуляризации [29].

3.2 Результаты

Для каждого метода был проведен поиск оптимальных параметров по сетке с помощью 5-блоковой перекрестной проверки. Оценка качества производилась с помощью коэффициента корреляции Мэтьюза mcc (Matthews correlation coefficient) - сбалансированной меры [30]. Полученные результаты приведены в таблице 1.

В таблице 2 указаны результаты оценки качества работы алгоритмов на тестовой выборке с помощью доли правильных ответов (accuracy) и F-меры (f1-score) [31] для каждого класса - уровня риска. На многоклассовой классификации accuracy не несет много информации, но, рассматриваяя ее совместно с f1-score, получается достаточно релеватная оценка. Жирным выделены лучшие показатели для каждого метода.

Таблица 1: Оптимальные параметры методов

	KNN	SVM	RF	LR
Skip-Gram	90	0.5	кол-во деревьев=300, глубина=20, мин. кол-во в ведущем узле=2, мин. кол-во в листовом узле=1	'newton-cg', c=0.5
FastText Self-study	90	0.25	кол-во деревьев=400, глубина=18, мин. кол-во в ведущем узле=2, мин. кол-во в листовом узле=1	'newton-cg', c=0.75
FastText	30	0.4	кол-во деревьев=200, глубина=16, мин. кол-во в ведущем узле=2, мин. кол-во в листовом узле=2	'newton-cg', c=1
BertEmbefings	40	0.25	кол-во деревьев=200, глубина=20, мин. кол-во в ведущем узле=2, мин. кол-во в листовом узле=1	'newton-cg', c=1

Таблица 2: Качество на тестовом наборе

		Skip-Gram	FastText Self-study	FastText	BertEmbedings
KNN	accuracy	0.58	0.57	0.54	**0.57**
	f1-score	0.12, 0.61, 0.68	0.11, 0.59, 0.68	0.15, 0.55, 0.64	**0.22, 0.56, 0.68**
SVM	accuracy	0.73	**0.74**	0.69	0.70
	f1-score	0.66, 0.7, 0.8	0.67, 0.72, 0.81	0.57, 0.68, 0.78	0.63, 0.68, 0.76
RF	accuracy	0.78	**0.79**	0.76	0.77
	f1-score	0.73, 0.77, 0.83	075, 0.77, 0.83	0.71, 0.74, 0.8	0.75, 0.76, 0.81
LR	accuracy	0.74	**0.75**	0.72	0.71
	f1-score	0.67, 0.7, 0.81	0.69, 0.72, 0.82	0.64, 0.69, 0.79	0.64, 0.69, 0.78

Модель FastText, самостоятельно обученная на тематических текстах, показала наиболее высокие результаты для 3 из 4 алгоритмов, уступив Word2Vec и BertEmbedings. Самообученный Word2Vec, показал схожие результаты, немного превзойдя FastText в метрическом алгоритме. Из этого можно сделать вывод, что Word2Vec реализует более точную близость векторов текстов, а FastText путем обработки опечаток и редких слов - содержательность.

Предобученные векторные представления показали в целом более низкие результаты, из чего можно сделать вывод, что при обработке медицинских текстов неэффективно использовать готовые подходы без модификаций.

Среди алгоритмов классификации лучшие результаты показал ансамблиевый RF, а худшие - метрический KNN и линейный SVM. Несмотря на достаточно высокую accuracy в методе опорных векторов, f1-score указывает на дисбаланс классов. Аналогично KNN подвержен сильному смещению к классу с большим количеством элементов. Из этого следует, что вектора пациентов с различными уровнями риска не сгруппированы в пространстве, несмотря на использование векторных представлений, поддерживающих семантически нагруженные линейные операции.

Вывод

Анализ литературы о применении методом машинного обучения в медицинской диагностике показал, что пока не существует единого подхода к разработке медицинских информационных систем. Разнородная структура текстов, неоднозначные сокращения и большое количество опечаток вызывают трудности обработки данных из электронных медицинских карт. Множество исследований и публикаций продолжают раскрывать перспективные методы решения задачи, отмечая ее значимость.

Изучение структуры предоставленной базы данных позволило определить наиболее значимые поля электронных медицинских карт. Результаты работы алгоритмов классификации показали, что выбор представления текстовых данных в качестве вложений оправдал себя. Лучшего результата удалось добиться с помощью метода FastText, обученного на собранных текстах. При этом использование предобученных моделей оказалось менее результативным, что еще раз подтверждает актуальность данной работы и специфичность обработки медицинских данных.

Рис. 8: Алгоритм работы.

Результаты работы четырех алгоритмов классификации помогли лучше понять распределение полученных многомерных данных и определить

слабые места реализованной модели. Перспективой для дальнейшего увеличения точности поставленной задачи является уточнение разметки записей, обработка несбалансированности и использование для классификации нейронных сетей (LSTM и сверточных).

Исходя из полученных результатов была составлена блок-схема алгоритма определения уровня риска, представленная на рисунке 8. Среднее время его работы для 100 пациентов - 21 секунда. Обработка такого количества запросов вручную заняла бы гораздо большее время.

Заключение

В ходе работы были выполнены следующие задачи:

1. Исследование структуры базы данных и анализ представленной в ней информации.

2. Изучение возможных подходов обработки данных и их применение на полученных признаках для создания представления пациента.

3. Обучение собственных моделей Word2Vec и FastText и оценка их эффективности.

4. Подбор параметров и анализ результатов 4 алгоритмов классификации.

5. Разработка программного кода [32], реализующего:

- сбор данных из вышеописанной базы данных,
- предобработку данных,
- векторизацию данных,
- классификацию полученного векторного представления.

Список литературы

[1] Медицинская нейроинформатика [Электронный ресурс]. URL: https://intuit.ru/studies/courses/1605/141/lecture/20589 (дата обращения: 20.12.20)

[2] Региональный акушерский мониторинг в Свердловской области - инновационный инструмент для снижения материнской и перинатальной смертности. Новые возможности дистанционной помощи / Н. О. Анкудинов [и др.] // Журнал телемедицины и электронного здравоохранения. 2015. № 1(1). С. 28-31.

[3] Зильбер Н. А., Анкудинов Н. О. Региональный акушерский мониторинг: инновационный инструмент управления кластером родовспоможения // Журнал телемедицины и электронного здравоохранения. 2019. Т. 5. № 1. С. 3-7.

[4] Гергет О. М. Модель и инструментальные средства анализа информационных процессов биологической системы мать-плод [Электронный ресурс]. URL: https://postgraduate.tusur.ru/system/file_copies/files/000/000/907/original/dissertation.pdf (дата обращения: 15.12.20)

[5] SF Medic [Электронный ресурс]. URL: https://www.sourcefuse.com/sf-medic (дата обращения: 18.12.20)

[6] How SF Medic Provides Real-Time Clinical Decision Support Using AWS Machine Learning Services [Электронный ресурс]. URL: https://aws.amazon.com/ru/blogs/apn/how-sf-medic-provides-real-time-clinical-decision-support-using-aws (дата обращения: 18.12.20)

[7] Choi E., Schuetz A. , Stewart W. F., Sun J. Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction [Электронный ресурс]. URL: https://arxiv.org/ftp/arxiv/papers/1602/1602.03686.pdf (дата обращения: 12.12.20)

[8] MySQL Documentation [Электронный ресурс]. URL: https://dev.mysql.com/doc/ (дата обращения: 21.10.20)

[9] Фиайли К. SQL. Москва: ДМК Пресс. 456

[10] Перинатальный риск, баллы [Электронный ресурс]. URL: http://www.chelsma.ru/files/misc/perinatalnyefaktoryriska.pdf (дата обращения: 10.02.21)

[11] Natural Language Processing (NLP) for Machine Learning [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/natural-language-processing-nlp-for-machine-learning-d44498845d5b/ (дата обращения: 21.01.21)

[12] Векторное представление слов [Электронный ресурс]. URL:https://neerc.ifmo.ru/wiki/index.php?title=%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%B5%D0%B4%D1%81%D1%82%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81%D0%BB%D0%BE%D0%B2 (дата обращения: 25.02.21)

[13] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space [Электронный ресурс]. URL: https://arxiv.org/pdf/1301.3781v3.pdf (дата обращения: 25.02.21)

[14] Документация Gensim Word2Vec [Электронный ресурс]. URL: https://radimrehurek.com/gensim/models/word2vec.html (дата обращения: 20.03.21)

[15] Документация FastText [Электронный ресурс]. URL: https://fasttext.cc/ (дата обращения: 22.03.21)

[16] Документация Gensim FastText [Электронный ресурс]. URL: https://radimrehurek.com/gensim/models/fasttext.html (дата обращения: 22.03.21)

[17] Devlin J., Chang M., Lee K., Toutanova K.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ре-

сурс]. URL: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 10.04.21)

[18] BERT in DeepPavlov [Электронный ресурс]. URL: http://docs.deeppavlov.ai/en/master/features/models/bert.html (дата обращения: 11.04.21)

[19] Дьяконов А.Г. Python: категориальные признаки [Электронный ресурс]. URL: https://dyakonov.org/2016/08/03/python (дата обращения: 03.03.21)

[20] Воронцов К.В. Математические методы обучения по прецедентам [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 28.03.21)

[21] Nearest Neighbors // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/neighbors.html#classification (дата обращения: 01.04.21)

[22] KNeighborsClassifier // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата обращения: 01.04.21)

[23] SVM // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.ru/1-4-support-vector-machines/ (дата обращения: 02.04.21)

[24] SVC // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC (дата обращения: 02.04.21)

[25] LinearSVC // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC (дата обращения: 02.04.21)

[26] Ensemble methods // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/ensemble.html#forest (дата обращения: 07.04.21)

[27] Воронцов К. В. Логические алгоритмы классификации [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/images/3/3e/Voron-ML-Logic.pdf (дата обращения: 06.04.21)

[28] RandomForestClassifier // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier (дата обращения: 06.04.21)

[29] LogisticRegression // scikit-learn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (дата обращения: 15.04.21)

[30] Matthews corrcoef // scikit-learn [Электронный ресурс]. URL:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html (дата обращения: 10.04.21)

[31] Оценка качества в задачах классификации и регрессии [Электронный ресурс]. URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B0%D1%85_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D0%B8_%D0%B8_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%B8 (дата обращения: 14.04.21)

[32] GitHub репозиторий [Электронный ресурс]. URL: https://github.com/SvetaGolovina/RiskClassification