АКАДЕМИЯ ГОСУДАРСТВЕННОГО УПРАВЛЕНИЯ

ПРИ ПРЕЗИДЕНТЕ АЗЕРБАЙДЖАНСКОЙ РЕСПУБЛИКИ

Факультет: *Административное управление*

Кафедра: *Информационная технология в государственном управлении*

Специальность: *Компьютерные науки*

Группа: *K212*

Отдел: *Очное*

К У Р С О В А Я Р А Б О Т А

НА ТЕМУ:

Характеристики и классификация инструментов Data Mining

Студент: Фатуллаев Айдын Малик оглу

Преподаватель: д.ф.э, доц Аскерова Б.

Заведующий кафедры: к.э.н, доц. Э. А. Абасов

БАКУ – 2024

Содержание

Введение в тему

Допустим есть большая база данных, в которой хранятся очень много записей. Возможно, что в этой “горе” информации среди обычных, не представляющих никому интереса, записей таятся и полезные. Естественно, специалисту самостоятельно просмотреть все записи не представится возможным, и он воспользуется помощью компьютера, а именно инструментами Data Mining.

Data Mining — это один из наиболее полезных методов, который помогает предпринимателям, исследователям и частным лицам извлекать ценную информацию из огромных наборов данных. Интеллектуальный анализ данных (Data Mining) также называется обнаружением знаний в базе данных. Процесс обнаружения знаний включает в себя очистку данных, интеграцию данных, выбор данных, преобразование данных, интеллектуальный анализ данных, оценку шаблонов и представление знаний. Прежде всего, он превращает необработанные данные в полезную информацию.

Вся суть Data Mining-а заключается в использовании различных инструментов. Инструменты Data Mining — это наборы программных средств, с помощью которых выполняется подготовка данных и обеспечиваются алгоритмы их интеллектуального анализа, а также осуществляются процессы машинного обучения. Эти программные средства и алгоритмы, используются для обнаружения закономерностей, шаблонов и взаимосвязей в больших объемах данных. Инструменты Data Mining позволяют извлекать ценные знания из данных, которые могут быть использованы для принятия бизнес-решений, оптимизации процессов и многих других целей.

В курсовой работе приведены несколько популярных инструментов Data Mining, отражены характеристики и классификация инструментов, а также отмечены примеры и цели применения инструментов Data Mining.

Что такое Data Mining?

Понятия о Data Mining

Что же такое Data Mining? Это процесс извлечения информации для выявления закономерностей, тенденций и полезных данных, которые позволяют бизнесу принимать решения на основе данных из огромных наборов данных, называется интеллектуальным анализом данных (Data Mining). Data Mining — это процесс исследования скрытых шаблонов информации с различных точек зрения для категоризации полезных данных. Ключевые шаги в процессе Data Mining включают выбор и предварительную обработку данных, применение алгоритмов анализа и интерпретацию полученных результатов.

Интеллектуальный анализ данных аналогичен науке о данных, выполняемой человеком в конкретной ситуации, на определенном наборе данных и с определенной целью. Этот процесс включает в себя различные типы услуг, такие как интеллектуальный анализ текста, веб-майнинг, аудио- и видеоинжиниринг, интеллектуальный анализ графических данных и интеллектуальный анализ социальных сетей. Это делается с помощью программного обеспечения, которое является простым или очень специфичным. Аутсорсинг интеллектуального анализа данных позволяет выполнить всю работу быстрее и с низкими эксплуатационными расходами. Специализированные фирмы также могут использовать новые технологии для сбора данных, которые невозможно найти вручную. На различных платформах доступны тонны информации, но доступных знаний очень мало. Самая большая проблема — проанализировать данные для извлечения важной информации, которую можно использовать для решения проблемы или для развития компании. Существует множество мощных инструментов и методов, позволяющих анализировать данные и находить на их основе более глубокое понимание.[8]

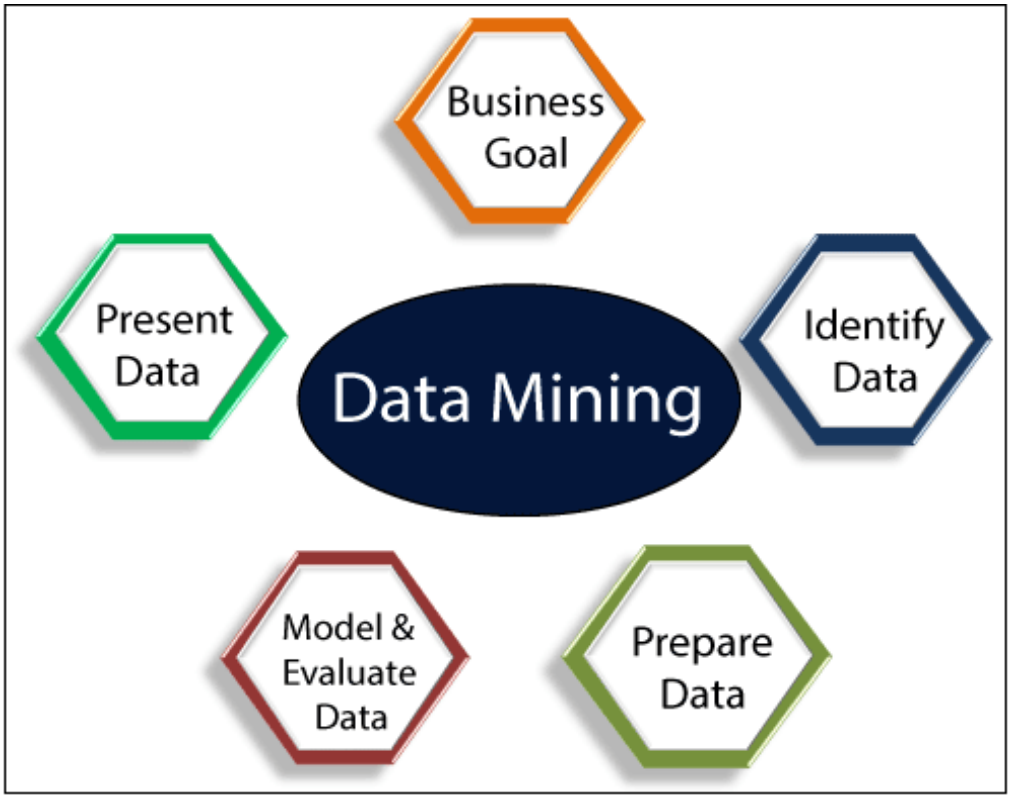

Рис. 1: Операции, проделываемые с данными в Data Mining

История интеллектуального анализа данных

В 1990-х годах был введен термин «Интеллектуальный анализ данных», но интеллектуальный анализ данных — это эволюция сектора с обширной историей.

Ранние методы выявления закономерностей в данных включают теорему Байеса (1700-е годы) и эволюцию регрессии (1800-е годы). Развитие и растущая мощь информатики способствовали увеличению сбора, хранения и обработки данных. Явное практическое исследование данных постепенно совершенствовалось за счет косвенной автоматической обработки данных и других открытий в области информатики, таких как нейронные сети, кластеризация, генетические алгоритмы (1950-е годы), деревья решений (1960-е годы) и вспомогательные векторные машины (1990-е годы).

Истоки интеллектуального анализа данных восходят к трем семейным линиям: классическая статистика, искусственный интеллект и машинное обучение.

Классическая статистика является основой большинства технологий, на которых построен интеллектуальный анализ данных, таких как регрессионный анализ, стандартное отклонение, стандартное распределение, стандартное отклонение, дискриминационный анализ, кластерный анализ и доверительные интервалы. Все они используются для анализа данных и подключения к данным.

ИИ или Искусственный интеллект основан на эвристике, а не на статистике. Он пытается применить человеческое мышление, например обработку данных, к статистическим задачам. Определенная концепция искусственного интеллекта была принята в некоторых высокопроизводительных коммерческих продуктах, таких как модули оптимизации запросов для системы управления реляционными базами данных (СУБД).

Машинное обучение — это сочетание статистики и искусственного интеллекта. Его можно рассматривать как эволюцию ИИ, поскольку он сочетает в себе эвристику ИИ со сложным статистическим анализом. Машинное обучение пытается дать компьютерным программам возможность узнать об изучаемых данных, чтобы программы могли принимать четкое решение на основе характеристик исследуемых данных. Он использует статистику для основных концепций и добавляет дополнительные эвристики и алгоритмы искусственного интеллекта для достижения своей цели. [11]

Инструменты Data Mining

Классификация инструментов Data Mining

Инструменты Data Mining можно классифицировать по различным критериям, включая функциональность, поддерживаемые методы анализа данных, доступность, стоимость и т. д. Вот некоторые общие категории, на которые можно разделить инструменты Data Mining:

1. По типу алгоритмов
2. По способу работы
3. По масштабу
4. По типу использования

Такое разнообразие инструментов вызвано тем, что их много на рынке. Каждый инструмент хорош по-своему, и занимает свою позицию на том рынке, где есть на него спрос. Т.е. программные средства различаются между собой и каждый из них лучше в своей области.

Классификация инструментов по типу алгоритмов:

По типу алгоритмов инструменты может классифицировать так:

1. Кластеризация
2. Классификация
3. Регрессия
4. Ассоциативные правила
5. Аномалии/выбросы

Кластеризация

Метод k-средних: инструмент кластеризации, который разбивает данные на заранее определенное количество кластеров, минимизируя суммарное квадратичное отклонение точек от центров кластеров.

Основанная на плотности пространственная кластеризация для приложений с шумами (DBSCAN): этот алгоритм ищет плотные области в

пространстве данных, определяя кластеры как непрерывные участки высокой плотности.

Иерархическая кластеризация: инструмент, который строит иерархическую структуру кластеров, объединяя или разделяя их на основе меры близости между кластерами и объектами.

Классификация:

Метод случайного леса: этот метод строит множество деревьев решений во время обучения и объединяет их для получения более точных и устойчивых результатов.

Метод опорных векторов: в этом методе пытаются найти гиперплоскость, которая наилучшим образом разделяет классы в пространстве признаков.

Байесовский классификатор: использует теорему Байеса для классификации объектов на основе их признаков.

Регрессия:

Линейная регрессия: простой метод, который пытается установить линейную зависимость между независимыми переменными и зависимой переменной.

Гребневая регрессия: вариант линейной регрессии, который также учитывает штраф за большие коэффициенты модели.

Лассо-регрессия: ещё один вариант линейной регрессии, который обычно применяется для отбора признаков и снижения переобучения.

Ассоциативные правила:

Алгоритм Apriori: этот алгоритм используется для нахождения часто встречающихся комбинаций элементов в транзакционных данных.

Frequent Pattern Growth (выращивание популярных часто встречающихся предметных наборов): более эффективный алгоритм по сравнению с Apriori для поиска частых наборов элементов в транзакционных данных.

Аномалии и выбросы:

Изоляционный лес: этот метод использует деревья решений для выделения аномальных объектов путем их изоляции в отдельные ветви деревьев.

Локальный коэффициент выброса: определяет аномальные объекты, анализируя плотность окрестности каждого объекта по сравнению с плотностью его соседей.[7][9]

Классификация инструментов по способу работы

Интерактивные инструменты Data Mining предоставляют пользователю возможность взаимодействия с данными в режиме реального времени. Пользователь может настраивать параметры анализа, визуализировать результаты, проводить эксперименты и манипулировать данными непосредственно в процессе работы. К ним относятся такие инструменты как: RapidMiner, Tableau, KNIME и другие.

Пакетные инструменты Data Mining выполняют анализ данных в автономном режиме без участия пользователя после начальной настройки. Пользователь определяет параметры анализа заранее, запускает процесс и ожидает завершения для просмотра результатов. К ним относятся такие инструменты как: Weka, инструменты, написанные на языках Python и R. [7]

Классификация инструментов по масштабу работы

Локальные инструменты: работают на отдельном компьютере или локальном сервере. Они обрабатывают данные в пределах одной машины и часто используют ресурсы (память, процессор) только этой машины. К примеру: Weka, Orange, DataMelt и другие.

Распределенные инструменты Data Mining могут параллельно выполнять вычисления на нескольких узлах или серверах для обработки больших объемов данных. Они могут масштабироваться горизонтально, увеличивая количество узлов для обработки данных. Примеры распределенных инструментов: SAS, SPSS, инструменты MATLAB и библиотеки Python и R. [7]

Классификация инструментов по типу использования

Открытые и бесплатные: Инструменты с открытым исходным кодом или бесплатные для использования. RapidMiner, Weka, Orange являются бесплатными.

Проприетарные: Инструменты, за использование которых требуется платить лицензионные сборы. IBM SPSS Modeler, SAS Enterprise Miner, KNIME Analytics Platform требуют оплаты за пользование. [7]

Характеристики инструментов Data Mining

Характеристики инструментов Data Mining могут быть разнообразными и зависят от их функциональности, возможностей, применимости, типа лицензии и т. д.

Вот некоторые ключевые характеристики, которые можно учесть при выборе инструмента Data Mining: Функциональность, Интерфейс, Поддержка данных, Масштабируемость, Производительность, Доступность, Совместимость, Поддержка и сообщество:

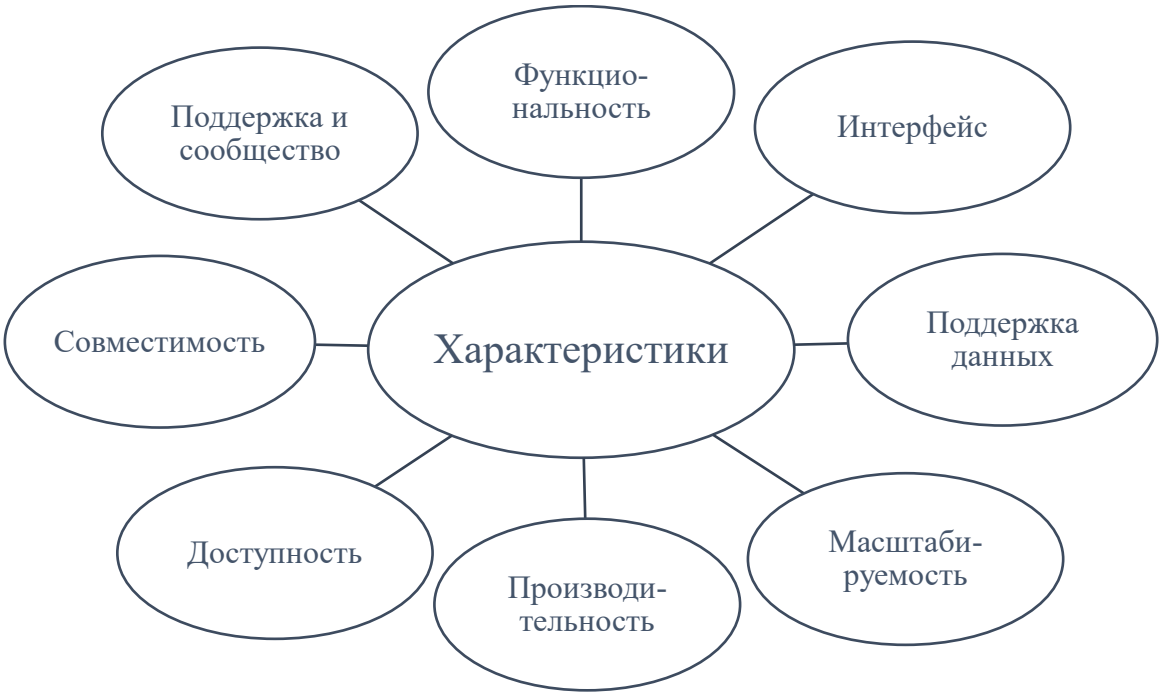

Рис. 2: Характеристики инструментов Data Mining

При выборе какого-либо инструмента Data Mining надо пройтись сперва по его характеристикам. Убедитесь, что все характеристики соответствуют вашим требованиям. Вы можете рассматривать несколько инструментов, отвечая на вопросы ниже, и определить лучший вариант для ваших дел и бизнеса.

1. **Функциональность:** какие алгоритмы и методы анализа данных поддерживаются инструментом? Он специализируется на кластеризации, классификации, регрессии, ассоциативных правилах или других типах анализа данных?
2. **Интерфейс:** имеет ли инструмент графический пользовательский интерфейс (GUI), командную строку или API для программирования? Какой уровень удобства использования предоставляется?
3. **Поддержка данных:** Какие типы данных поддерживаются? Могут ли инструменты работать с структурированными или неструктурированными данными? Поддерживают ли они разные источники данных?
4. **Масштабируемость:** масштабируются ли вычисления на большие объемы данных? Поддерживается ли распределенная обработка для работы с большими кластерами данных?
5. **Производительность:** Какова скорость выполнения анализа данных? Могут ли инструменты оптимизировать вычисления для повышения производительности?
6. **Доступность:** является ли инструмент с открытым исходным кодом или проприетарным? Какова стоимость использования инструмента?
7. **Совместимость:** Совместимы ли инструменты с другими инструментами и платформами? Могут ли они интегрироваться с существующими системами?
8. **Поддержка и сообщество:** Каков уровень поддержки со стороны разработчиков или сообщества пользователей? Есть ли документация, учебные материалы, форумы поддержки и т. д.?

Примеры инструментов Data Mining, у которых есть различные характеристики, включают в себя Weka, RapidMiner, KNIME, Python с библиотеками Pandas и Scikit-learn, R с пакетами для анализа данных, SAS Enterprise Miner, IBM SPSS Modeler и другие. Выбор конкретного инструмента зависит от требований проекта, опыта пользователей и доступных ресурсов.

Целью инструментов интеллектуального анализа данных является обнаружение закономерностей/тенденций/группировок среди больших наборов данных и преобразование данных в более точную информацию.

Рынок инструментов Data Mining процветает: согласно последнему отчету ReportLinker, к 2023 году объем продаж рынка превысит 1 миллиард долларов по сравнению с 591 миллионом долларов в 2018 году.

Orange Data Mining

Orange — идеальный пакет программного обеспечения для машинного обучения и интеллектуального анализа данных. Он поддерживает визуализацию и представляет собой программное обеспечение на основе компонентов, написанных на вычислительном языке Python.

Поскольку это программное обеспечение основано на компонентах, компоненты Orange называются «виджетами». Эти виджеты варьируются от предварительной обработки и визуализации данных до оценки алгоритмов и прогнозного моделирования.

Виджеты предоставляют функциональные возможности, такие как:

1. Отображение таблицы данных и возможность выбора функций
2. Чтение данных
3. Обучение предикторов и сравнение алгоритмов обучения
4. Визуализация элементов данных и т. д.

Кроме того, Orange обеспечивает более интерактивную и приятную атмосферу для скучных аналитических инструментов. Работать довольно интересно.

Данные, поступающие в Orange, быстро форматируются по нужному шаблону, а перемещаемые виджеты можно легко перенести куда необходимо. Orange позволяет своим пользователям принимать более разумные решения за короткое время, быстро сравнивая и анализируя данные. Это хорошая визуализация данных с открытым исходным кодом, а также оценка, которая подходит как новичкам, так и профессионалам. Интеллектуальный анализ данных может выполняться с помощью визуального программирования или сценариев Python. Многие виды анализа возможны благодаря интерфейсу визуального программирования (перетаскивание, связанное с виджетами), и обычно поддерживаются многие визуальные инструменты, такие как гистограммы, диаграммы рассеяния, деревья, дендрограммы и тепловые карты. Обычно поддерживается значительное количество виджетов (более 100).

В инструменте есть компоненты машинного обучения, надстройки для биоинформатики и анализа текста, а также множество функций для анализа данных. Это также используется как библиотека Python.

Скрипты Python могут продолжать работать в окне терминала, интегрированной среде, такой как PyCharm и PythonWin, и оболочках PR, таких как iPython. Orange состоит из интерфейса холста, на котором пользователь размещает виджеты и создает рабочий процесс анализа данных. Виджет предлагает фундаментальные операции, например, чтение данных, отображение таблицы данных, выбор функций, обучение предикторов, сравнение алгоритмов обучения, визуализацию элементов данных и т. д. Orange работает в Windows, Mac OS X и различных операционных системах Linux. Orange поставляется с несколькими алгоритмами регрессии и классификации.

Orange может читать документы в собственном и других форматах данных. Orange специализируется на методах машинного обучения для классификации или контролируемого анализа данных. В классификации используются два типа объектов: обучающиеся и классификаторы. Учащиеся рассматривают данные на уровне класса и возвращают классификатор. Методы

регрессии очень похожи на классификацию в Orange, и оба предназначены для контролируемого интеллектуального анализа данных и требуют данных на уровне класса. Обучение ансамблей объединяет прогнозы отдельных моделей для повышения точности. Модель может быть либо основана на разных обучающих данных, либо использовать разных учащихся для одних и тех же наборов данных.[4][10]

SAS Data Mining

SAS означает систему статистического анализа. Это продукт создан для аналитики и управления данными. SAS может собирать данные, изменять их, управлять информацией из различных источников и анализировать статистику. Он предлагает графический интерфейс для нетехнических пользователей.

Датамайнер SAS позволяет пользователям анализировать большие данные и предоставлять точную информацию для своевременного принятия решений. SAS имеет архитектуру обработки распределенной памяти, которая хорошо масштабируется. Он подходит для интеллектуального анализа данных, оптимизации и анализа текста. [10]

Пользователи могут настраивать параметры моделей, выбирать и адаптировать алгоритмы анализа данных в соответствии с требованиями и особенностями своих данных и задач. SAS Data Mining обеспечивает высокий уровень безопасности и защиты данных, включая механизмы аутентификации, авторизации, шифрования и аудита.

SAS Data Mining интегрируется с другими продуктами SAS, такими как SAS Visual Analytics, SAS Viya, SAS Enterprise Miner и другими, что расширяет его возможности и обеспечивает единый рабочий процесс анализа данных.[1]

DataMelt

DataMelt — это среда вычислений и визуализации, предлагающая интерактивную структуру для анализа и визуализации данных. Он

предназначен в первую очередь для студентов, инженеров и ученых. Он также известен как DMelt.

DataMelt — многоплатформенная утилита, написанная на Java. Он может работать в любой операционной системе, совместимой с JVM (виртуальной машиной Java). Он состоит из научных и математических библиотек. Научные библиотеки используются для построения 2D/3D графиков. Математические библиотеки используются для генерации случайных чисел, алгоритмов, подбора кривых и т. д.

DataMelt можно использовать для анализа больших объемов данных, интеллектуального анализа данных и статистического анализа. Он широко используется в естественных науках, финансовых рынках и технике.[6][10]

Rattle

Ratte — это инструмент интеллектуального анализа данных, основанный на графическом интерфейсе. Он использует язык программирования статистики R. Rattle раскрывает статическую мощь R, предлагая важные функции интеллектуального анализа данных. Несмотря на то, что Rattle имеет комплексный и хорошо развитый пользовательский интерфейс, он имеет встроенную вкладку кода журнала, которая создает дублирующийся код для любой операции графического пользовательского интерфейса.

Набор данных, созданный Rattle, можно просматривать и редактировать. Rattle дает другому возможность просматривать код, использовать его для различных целей и расширять код без каких-либо ограничений.

Rattle — это аббревиатура от “R Analytical Tool To Learn Easily”. Rattle доступен в системах Windows, Mac и Linux.[5][10]

Rapid Miner

Rapid Miner — одна из самых популярных систем прогнозного анализа, созданная одноименной с Rapid Miner компанией. Он написан на языке программирования Java. Он предлагает интегрированную среду для

интеллектуального анализа текста, глубокого обучения, машинного обучения и прогнозного анализа.

Инструмент можно использовать для широкого спектра приложений, включая корпоративные приложения, коммерческие приложения, исследования, образование, обучение, разработку приложений и машинное обучение.

Rapid Miner предоставляет сервер как на месте, так и в публичной или частной облачной инфраструктуре. В основе лежит модель клиент/сервер. Быстрый майнер поставляется с платформами на основе шаблонов, которые обеспечивают быструю доставку с небольшим количеством ошибок (которые обычно ожидаются в процессе написания кода вручную).[10]

Заключение

В курсовой работе была разобрана тема инструментов Data Mining. Они играют важную роль в современной аналитике данных, предоставляя средства для извлечения ценной информации из больших объемов данных. Они обладают различными характеристиками, что позволяет пользователям выбирать инструменты в зависимости от их специфических потребностей и требований проекта.

Если ваш проект нуждается в переработке больших объемов данных, то вам следует выбирать инструменты Python или SAS Data Mining. Если вам требуются графики или другое визуальное представление данных, можно выбрать Orange или Rattle, у которого очень удобный интерфейс. Для машинного обучения подходят Rapid Miner, инструменты использующие библиотеки Python, Java и т.д.

Подводя итоги, можно сказать, что инструменты Data Mining играют ключевую роль в процессе преобразования данных в ценную информацию, что в конечном итоге способствует развитию бизнеса, науки и общества в целом.

Список использованной литературы:

1. https://data.korusconsulting.ru/platforms/sas/sas-enterprise-miner/
2. https://hsbi.hse.ru/articles/instrumenty-data-mining
3. https://iso.ru/ru/press-center/journal/2135.phtml
4. https://orangedatamining.com/
5. https://r4stats.com/articles/software-reviews/rattle/
6. https://ru.wikipedia.org/wiki/DataMelt
7. https://studfile.net/preview/5554364/page:85/
8. https://www.javatpoint.com/data-mining
9. https://www.javatpoint.com/data-mining-techniques
10. https://www.javatpoint.com/data-mining-tools
11. https://www.javatpoint.com/history-of-data-mining