Изображение выглядит как текст, письмо, снимок экрана, документ

Автоматически созданное описание

Изображение выглядит как текст, письмо, бумага, Шрифт

Автоматически созданное описание

План

1. Содержание
2. Введение
3. 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

В него входят следующие подразделы:

1.1. Описание предметной области

1.1.1. Значимость анализа данных в рассматриваемой области.

1.1.2. Основные типы данных (числовые, категориальные, временные ряды).

1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов).

1.2. Основные методы статистической обработки данных

1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия).

1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат).

1.2.3. Корреляционный анализ.

1.2.5. Примеры практического применения статистических методов.

1.3. Язык программирования R и его возможности для статистического анализа

1.3.1. Основные преимущества языка R.

1.3.2. Обзор ключевых библиотек (tidyverse, ggplot2, dplyr, caret, forecast).

1.3.3. Примеры использования R в обработке данных.

1.4. Программное обеспечение Glarus BI

1.4.1. Общая характеристика BI-системы.

1.4.2. Основные функциональные возможности (импорт данных, визуализация, анализ).

1.4.3. Примеры использования Glarus BI для анализа данных.

1.5. Сравнение возможностей R и Glarus BI

1.5.1. Где R удобнее (гибкость, мощность анализа, поддержка сложных моделей).

1.5.2. Где Glarus BI удобнее (интерактивные отчёты, простота использования).

1.5.3. Возможность совместного использования инструментов.

Содержание подразделов

1.1. Описание предметной области

В этом разделе студент описывает, какую область данных он анализирует, почему анализ данных важен для данной сферы, какие основные задачи решаются с помощью статистики. Рассматриваются ключевые типы данных, с которыми ведётся работа.

1.2. Основные методы статистической обработки данных

Этот подраздел посвящён статистическим методам, применяемым в исследовании. Описываются основные методы анализа, их применение в различных сферах и примеры использования.

1.3. Язык программирования R и его возможности

Раздел включает краткую характеристику R, его ключевые преимущества, библиотеки для анализа данных и примеры их использования.

1.4. Программное обеспечение Glarus BI

Рассматриваются основные функции Glarus BI, возможности автоматизированного анализа данных, интерактивная визуализация и примеры использования.

1.5. Сравнение возможностей R и Glarus BI

Анализируются сильные и слабые стороны R и Glarus BI, их возможности в контексте статистической обработки данных, преимущества их совместного использования.

Вывод по первому разделу

Первый раздел закладывает теоретическую основу исследования, объясняя выбор статистических методов и инструментов анализа данных. Он формирует понимание задач исследования и определяет инструменты, которые будут применяться в практической части работы.

Крч тут описать цели и задачи, актуальность, примененные методы в работе (то бишь метод опороных векторов), облизать R и Glarus BI (как-то), описание EDA методов (статистика, корреляция).

Оглавление

[ВВЕДЕНИЕ 9](#_Toc198046950)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 11](#_Toc198046951)

[Глава 1. Теоретические основы классификации музыкальных жанров (расширенные пояснения) 11](#_Toc198046952)

[• 1.1. Описание предметной области 11](#_Toc198046953)

[• 1.1.1. Значимость анализа данных в рассматриваемой области: Роль автоматической классификации музыки в современном цифровом мире. 11](#_Toc198046954)

[\* Раскрыть экономические аспекты: влияние на стриминговые сервисы, персонализированную рекламу. 11](#_Toc198046955)

[\* Раскрыть технологические аспекты: машинное обучение, Big Data, cloud computing. 11](#_Toc198046956)

[\* Привести примеры использования: Pandora, Spotify, Apple Music. Объяснить, как классификация помогает в рекомендательных системах, плейлистах, автоматической категоризации. 11](#_Toc198046957)

[• 1.1.2. Основные типы данных (числовые, категориальные, временные ряды): Аудио данные как совокупность временных рядов и их преобразование в числовые признаки. 11](#_Toc198046958)

[\* Типы данных в контексте музыкальных данных: 11](#_Toc198046959)

[\* Аудиоданные (временные ряды): Представляют собой изменение амплитуды звука во времени. Это сырой, необработанный формат. Нужно объяснить суть оцифровки звука: частота дискретизации, квантование. 11](#_Toc198046960)

[\* Числовые данные (аудио признаки): Извлекаются из аудио данных с помощью различных алгоритмов. Примеры (детализировать каждый): 11](#_Toc198046961)

[Темп (BPM):\* Количество ударов в минуту, отражает скорость музыки. 11](#_Toc198046962)

[Тональность:\* Основная тональность произведения (например, До мажор, Ля минор). 11](#_Toc198046963)

[Яркость (Brightness):\* Мера "блеска" или "резкости" звука, связанная с частотным составом. 11](#_Toc198046964)

[Громкость (Loudness):\* Субъективное восприятие силы звука. 11](#_Toc198046965)

[Спектральный центр (Spectral Centroid):\* Показывает, где в спектре сконцентрирована основная энергия звука. 12](#_Toc198046966)

[Спектральный разброс (Spectral Spread):\* Мера распределения энергии вокруг спектрального центра. 12](#_Toc198046967)

[MFCC (Mel-Frequency Cepstral Coefficients):\* Набор признаков, отражающих форму спектральной огибающей звука (очень важны для распознавания речи и музыки). 12](#_Toc198046968)

[\* Категориальные данные (жанры): Метки, присваиваемые музыкальным произведениям (рок, поп, джаз, классика и т.д.). 12](#_Toc198046969)

[\* Обязательно указать, что сырые аудиоданные сами по себе непригодны для SVM, необходимо извлечь признаки. 12](#_Toc198046970)

[• 1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов): Классификация как задача прогнозирования принадлежности к определенному жанру. 12](#_Toc198046971)

[\* Прогнозирование: В данном контексте – прогнозирование жанра музыкального произведения на основе его аудио признаков. Подчеркнуть, что SVM, как алгоритм машинного обучения, является прогностическим инструментом. 12](#_Toc198046972)

[\* Выявление закономерностей: Цель - найти статистические закономерности, связывающие аудио признаки и жанры. Это позволит понять, какие характеристики звука типичны для каждого жанра. 12](#_Toc198046973)

[\* Оптимизация процессов: В данном контексте оптимизация может касаться: 12](#_Toc198046974)

[Оптимизации извлечения признаков:\* Выбор наиболее информативных признаков. 12](#_Toc198046975)

[Оптимизации параметров SVM:\* Настройка параметров ядра и регуляризации для достижения максимальной точности классификации. 12](#_Toc198046976)

[• 1.2. Основные методы статистической обработки данных 12](#_Toc198046977)

[• 1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия): Применение для анализа аудио признаков. 12](#_Toc198046978)

[\* Описание: Дать определение описательной статистики – это методы, используемые для обобщения и представления данных в понятной форме. 13](#_Toc198046979)

[\* Применение к аудио признакам: 13](#_Toc198046980)

[Среднее (Mean):\* Среднее значение аудио признака (например, среднего темпа всех песен в жанре рок). Позволяет оценить типичное значение признака для данного жанра. 13](#_Toc198046981)

[Медиана (Median):\* Значение, разделяющее упорядоченный набор данных пополам. Менее чувствительна к выбросам, чем среднее. 13](#_Toc198046982)

[Мода (Mode):\* Наиболее часто встречающееся значение в наборе данных. Показывает наиболее типичное значение признака. 13](#_Toc198046983)

[Дисперсия 13](#_Toc198046984)

[ChatGPT4 | Midjourney, [13.05.2025 14:13] 13](#_Toc198046985)

[(Variance):\* Мера разброса значений вокруг среднего. Показывает, насколько сильно значения признака варьируются в данном жанре. 13](#_Toc198046986)

[Стандартное отклонение (Standard Deviation):\* Квадратный корень из дисперсии. Более интерпретируемая мера разброса, чем дисперсия, т.к. выражается в тех же единицах измерения, что и сам признак. 13](#_Toc198046987)

[\* Пример: "Средний темп песен в жанре рок составляет 120 BPM, со стандартным отклонением 15 BPM, что говорит о значительной вариативности темпа в данном жанре." 13](#_Toc198046988)

[• 1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат): Использование для оценки значимости различий между группами признаков. 13](#_Toc198046989)

[\* Описание: Дать определение методов проверки гипотез – это статистические процедуры, используемые для проверки утверждений о генеральной совокупности на основе выборочных данных. 13](#_Toc198046990)

[\* Применение к аудио признакам: 13](#_Toc198046991)

[t-тест (Student's t-test):\* Используется для сравнения средних значений двух групп. Например, для проверки гипотезы о том, что средний темп песен в жанре рок статистически значимо отличается от среднего темпа песен в жанре поп. 13](#_Toc198046992)

[ANOVA (Analysis of Variance):\* Используется для сравнения средних значений трех и более групп. Например, для проверки гипотезы о том, что средний темп песен статистически значимо различается между жанрами рок, поп и джаз. 14](#_Toc198046993)

[Хи-квадрат (Chi-square test):\* Используется для анализа категориальных данных. Например, для проверки гипотезы о том, что наличие определенного инструмента (например, гитары) статистически связано с определенным жанром. 14](#_Toc198046994)

[• 1.2.3. Корреляционный анализ: Оценка взаимосвязи между различными аудио признаками. 14](#_Toc198046995)

[\* Описание: Дать определение корреляционного анализа – это статистические методы, используемые для измерения степени линейной взаимосвязи между двумя или более переменными. 14](#_Toc198046996)

[\* Применение к аудио признакам: 14](#_Toc198046997)

[\* Например, может существовать корреляция между яркостью звука и уровнем энергии. Это может быть полезно при отборе признаков для классификации. 14](#_Toc198046998)

[\* Используется для выявления избыточных признаков (признаков, которые сильно коррелируют друг с другом). В этом случае можно исключить один из коррелирующих признаков, чтобы упростить модель и избежать мультиколлинеарности. 14](#_Toc198046999)

[\* Указать, что корреляция не подразумевает причинно-следственную связь. 14](#_Toc198047000)

[Важные замечания: 14](#_Toc198047001)

[• Для каждой статистической величины и метода приведите примеры применения конкретно к задачам классификации музыкальных жанров. 14](#_Toc198047002)

[• Не нужно углубляться в сложные математические детали, достаточно дать понятные определения и примеры. 14](#_Toc198047003)

[• Подчеркните, как эти методы будут использоваться в практической части вашей работы. 15](#_Toc198047004)

[• Обязательно указывайте ссылки на источники информации. 15](#_Toc198047005)

[Это расширенное описание позволит вам более подробно и качественно раскрыть теоретическую часть вашей курсовой работы. Удачи! 15](#_Toc198047006)

# ВВЕДЕНИЕ

Музыка является неотъемлемой частью культуры и играет важную роль в нашей жизни. С развитием цифровых технологий и стриминговых сервисов объемы доступной музыки растут многократно. Это создает потребность в эффективных методах автоматической организации и анализа музыкальных данных. Одним из ключевых аспектов является классификация музыкальных произведений по жанрам.

Автоматическая классификация музыкальных жанров – это задача машинного обучения, заключающаяся в определении жанра музыкального произведения на основе его аудио характеристик. Традиционные методы ручной классификации являются трудоемкими, субъективными и не масштабируемыми для обработки больших объемов данных. В связи с этим разработка эффективных и точных автоматизированных методов классификации музыкальных жанров представляет собой актуальную и важную задачу.

Настоящая курсовая работа посвящена разработке программы классификации музыкальных произведений по жанрам с использованием алгоритмов метода опорных векторов (SVM). Метод опорных векторов является мощным инструментом машинного обучения, который обладает высокой обобщающей способностью и эффективен в задачах классификации, в том числе для нелинейно разделимых данных.

Цель данной работы ­- разработка программы классификации музыкальных произведений по жанрам на основе алгоритмов метода опорных векторов (SVM), способной автоматически определять жанр музыкального произведения по его аудио характеристикам.

Для достижения поставленной цели необходимо решить следующие задачи:

* Изучить предметную область и провести анализ существующих методов классификации музыкальных жанров, выявив их достоинства и недостатки.
* Выполнить предварительную обработку данных, получив общее представление о распределении признаков, а также выявить возможные зависимости, аномалии и особенности структуры выборки.
* Провести обзор алгоритмов метода опорных векторов (SVM), изучив теоретические основы, различные типы ядер, методы оптимизации и параметры регуляризации.
* Разработать модель классификации, определив набор аудио характеристик (признаков), используемых для классификации, разработав архитектуру модели SVM и обучив ее на размеченном наборе данных.
* Реализовать программное обеспечение, реализующее разработанную модель классификации музыкальных жанров на языке R.
* Оценить эффективность разработанной программы, проведя тестирование на контрольном наборе данных и оценив ее точность, полноту и другие метрики качества.
* Визуализировать данные с помощью графики, диаграммы, дашборды, используя язык программирования R и возможности аналитической системы Glarus BI.

Объектом исследования являются синтетические данные музыкальных характеристик, описывающих уровни шума, ритма и других параметров для жанров музыки (рок, хип-хоп и поп). Предметом - метод опорных векторов (SVM) и его применение для классификации музыкальных жанров.

Практическая реализация выполняется в программной среде R с использованием различных библиотеки и инструментов для анализа и визуализации данных. Также уделяется внимание автоматизации анализа и формированию отчётов с помощью R Markdown, что позволяет создать удобный и расширяемый аналитический процесс.

# ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

# Описание предметной области

# 1.1.1. Значимость анализа данных в области музыкального бизнеса

В современном цифровом мире, где объем музыкального контента растет экспоненциально, классификация музыки играет критически важную роль. Она позволяет эффективно организовывать, искать и рекомендовать музыку, формируя пользовательский опыт и принося значительную выгоду различным участникам рынка.

Методы машинного обучения используются различными компаниями в сфере музыки. Стриминговые сервисы полагаются на классификацию музыки для организации огромных библиотек, улучшения поиска и, самое главное, для создания персонализированных рекомендаций и плейлистов. Чем точнее и релевантнее рекомендации, тем выше вовлеченность пользователей, дольше время, проведенное на платформе, и, как следствие, выше выручка от подписок и рекламы.

Также классификация музыкального контента помогает решать задачи, такие как:

* повышение удержания пользователей: персонализированные плейлисты, допустим, напрямую влияют на удержание пользователей, снижая отток и стимулируя продление подписок;
* оптимизация лицензионных соглашений: классификация музыки позволяет более точно определять жанры и стили, что важно при заключении лицензионных соглашений с правообладателями;
* монетизация контента: для независимых музыкантов и лейблов классификация предоставляет возможность эффективнее монетизировать свою музыку. Наиболее продвинутая модель классификации позволяет их трекам быть обнаруженными целевой аудиторией, попадать в релевантные плейлисты и, как следствие, генерировать больше прослушиваний и доходов.
* экономия ресурсов: модель классификации значительно экономит время и ресурсы по сравнению с ручной категоризацией. Это особенно важно для платформ с огромными музыкальными библиотеками, где ручная классификация получилась бы значительно дороже.

# 1.1.2. Основные типы данных (числовые, категориальные, временные ряды) Типы данных в контексте музыкальных данных:

* Аудиоданные (временные ряды): представляют собой изменение амплитуды звука во времени. Это сырой, необработанный формат. Нужно объяснить суть оцифровки звука: частота дискретизации, квантование. Однако в данной курсовой работе не используется).
* Числовые данные (аудио признаки): Извлекаются из аудио данных с помощью различных алгоритмов. В данной работе применяются следующие характеристики:
  + Темп (BPM - Количество ударов в минуту) отражает скорость музыки.
  + RMS Energy (Root Mean Square - Среднеквадратичная энергия) показывает общую громкость сигнала. Анализ должен определять среднее значение RMS Energy для трека и учитывать его влияние на восприятие музыки. Например, треки с высокой RMS Energy часто кажутся более энергичными и агрессивными, а треки с низкой RMS Energy - более тихими и спокойными.
  + Zero Crossing Rate (Сигнальный шум) демонстрирует частоту, с которой сигнал пересекает нулевую линию. Более высокие значения ZCR обычно указывают на более "шумный" или высокочастотный сигнал, который может быть связан с определенными инструментами (например, перкуссия) или жанрами (например, электронная музыка).
  + Инструментальность показывает вероятность того, что в треке отсутствует вокал. Значение 1.0 означает, что трек полностью инструментальный, а значение 0.0 - что в треке преобладает вокал.
  + Фактически, параметр "Вокал" обратно пропорционален "Инструментальности". Анализ может быть таким же, как и у "Инструментальности", или можно использовать этот параметр для уточнения классификации, если есть сомнения между жанрами, где вокал является определяющим (например, поп против инструментального хип-хопа).
* Категориальные данные (жанры): Метки, присваиваемые музыкальным произведениям (рок, поп, хип-хоп). Преобразуются в кодировку от 0 до 2 для определения класса прогнозирования.

### **1.1.3. Ключевые задачи статистического анализа (прогнозирование, выявление закономерностей, оптимизация процессов)**

К основным задачам статистического анализа в анализе музыкальных данных относятся:

* Прогнозирование: В данном контексте – прогнозирование жанра музыкального произведения на основе его аудио признаков. Подчеркнуть, что SVM, как алгоритм машинного обучения, является прогностическим инструментом.
* Выявление закономерностей: Цель - найти статистические закономерности, связывающие аудио признаки и жанры. Это позволит понять, какие характеристики звука типичны для каждого жанра.
* Оптимизации извлечения признаков: Выбор наиболее информативных признаков и оптимизация параметров SVM:\* Настройка параметров ядра и регуляризации для достижения максимальной точности классификации.

## **1.2. Основные методы статистической обработки данных**

**1.2.1. Описательная статистика (среднее, медиана, мода, дисперсия): Применение для анализа аудио признаков.**

\* Описание: Дать определение описательной статистики – это методы, используемые для обобщения и представления данных в понятной форме.

\* Применение к аудио признакам:

Среднее (Mean):\* Среднее значение аудио признака (например, среднего темпа всех песен в жанре рок). Позволяет оценить типичное значение признака для данного жанра.

Медиана (Median):\* Значение, разделяющее упорядоченный набор данных пополам. Менее чувствительна к выбросам, чем среднее.

Мода (Mode):\* Наиболее часто встречающееся значение в наборе данных. Показывает наиболее типичное значение признака.

Дисперсия

ChatGPT4 | Midjourney, [13.05.2025 14:13]

(Variance):\* Мера разброса значений вокруг среднего. Показывает, насколько сильно значения признака варьируются в данном жанре.

Стандартное отклонение (Standard Deviation):\* Квадратный корень из дисперсии. Более интерпретируемая мера разброса, чем дисперсия, т.к. выражается в тех же единицах измерения, что и сам признак.

\* Пример: "Средний темп песен в жанре рок составляет 120 BPM, со стандартным отклонением 15 BPM, что говорит о значительной вариативности темпа в данном жанре."

• 1.2.2. Методы проверки гипотез (t-тест, ANOVA, хи-квадрат): Использование для оценки значимости различий между группами признаков.

\* Описание: Дать определение методов проверки гипотез – это статистические процедуры, используемые для проверки утверждений о генеральной совокупности на основе выборочных данных.

\* Применение к аудио признакам:

t-тест (Student's t-test):\* Используется для сравнения средних значений двух групп. Например, для проверки гипотезы о том, что средний темп песен в жанре рок статистически значимо отличается от среднего темпа песен в жанре поп.

ANOVA (Analysis of Variance):\* Используется для сравнения средних значений трех и более групп. Например, для проверки гипотезы о том, что средний темп песен статистически значимо различается между жанрами рок, поп и джаз.

Хи-квадрат (Chi-square test):\* Используется для анализа категориальных данных. Например, для проверки гипотезы о том, что наличие определенного инструмента (например, гитары) статистически связано с определенным жанром.

• 1.2.3. Корреляционный анализ: Оценка взаимосвязи между различными аудио признаками.

\* Описание: Дать определение корреляционного анализа – это статистические методы, используемые для измерения степени линейной взаимосвязи между двумя или более переменными.

\* Применение к аудио признакам:

\* Например, может существовать корреляция между яркостью звука и уровнем энергии. Это может быть полезно при отборе признаков для классификации.

\* Используется для выявления избыточных признаков (признаков, которые сильно коррелируют друг с другом). В этом случае можно исключить один из коррелирующих признаков, чтобы упростить модель и избежать мультиколлинеарности.

\* Указать, что корреляция не подразумевает причинно-следственную связь.

Важные замечания:

• Для каждой статистической величины и метода приведите примеры применения конкретно к задачам классификации музыкальных жанров.

• Не нужно углубляться в сложные математические детали, достаточно дать понятные определения и примеры.

• Подчеркните, как эти методы будут использоваться в практической части вашей работы.

• Обязательно указывайте ссылки на источники информации.

Это расширенное описание позволит вам более подробно и качественно раскрыть теоретическую часть вашей курсовой работы. Удачи!